UNCLASSIFIED

REPORT DOCUMENTATION PAGE

AD-A197 496

1b. RESTRICTIVE MARKINGS
NA

DTIC FILE COPY

3. DISTRIBUTION / AVAILABILITY OF REPORT
Distribution Unlimited; Approved for Public Release

NA

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE
NA

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
INDU/DC/GMH/TR-39

5. MONITORING ORGANIZATION REPORT NUMBER(S)
NA

DTIC ELECTE AUG 16 1988 S D CE

6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Indiana University	NA	ONR

6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
Department of Chemistry Bloomington, IN 47405	800 N. Quincy Street Arlington, VA 22217

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
		Contract N00014-86-K-0366

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO. R&T Code 4134006	WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
Detection of Subpopulations in Near-Infrared Reflectance Analysis

12. PERSONAL AUTHOR(S)
Robert A. Lodder and Gary M. Hieftje

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	11 July 1988	65

16. SUPPLEMENTARY NOTATION
Accepted for publication in Applied Spectroscopy

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Near Infrared Reflectance Analysis, Chemometrics

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In typical near-infrared multivariate statistical analyses, samples with similar spectra produce points that cluster in a certain region of spectral hyperspace. These clusters can vary significantly in shape and size due to variations in sample packings, particle-size distributions, component concentrations, and drift with time. These factors, when combined with discriminant analysis using simple distance metrics, produce a test in which a result that places a particular point inside a particular cluster does not necessarily mean that the point is actually a member of the cluster. Instead, the point may be a member of a new, slightly different cluster that overlaps the first. A new cluster can be created by factors like low-level contamination or instrumental drift. An extension added to part of the BEAST (Bootstrap Error-Adjusted Single-sample Technique) can be used to set nonparametric probability-density contours inside spectral clusters as well as outside, and when multiple points begin to appear in a certain region of cluster-hyperspace the perturbation of these density contours can be detected at an assigned significance level. The detection of false samples both within and beyond 3 SDs of the center of the training set is possible with this method. This procedure is shown to be effective for contaminant levels of a few hundred ppm in an over-the-counter drug capsule, and is shown to function with as few as one or two wavelengths, suggesting its application to very simple process sensors.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Distribution Unlimited

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Gary M. Hieftje	(812) 335-2189	

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

OFFICE OF NAVAL RESEARCH

Contract N14-86-K-0366

R&T Code 4134006

TECHNICAL REPORT NO. 39

DETECTION OF SUBPOPULATIONS IN NEAR-INFRARED REFLECTANCE ANALYSIS

by

Robert A. Lodder and Gary M. Hieftje

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

Prepared for Publication

in

APPLIED SPECTROSCOPY

Indiana University
Department of Chemistry
Bloomington, Indiana 47405

11 July 1988

QUALITY INSPECTED 2

Reproduction in whole or in part is permitted for any purpose of the United States Government

This document has been approved for public release and sale; its distribution is unlimited

The uses of near-infrared spectrometry have increased rapidly since the introduction of multiple linear regression and other pattern-recognition techniques to near-IR spectral data analysis[1,2]. Quantitative analysis of mixtures in the near-IR region has proven to be a powerful method of examining routine samples (samples whose basic composition is known). Qualitative applications of near-IR spectrometry have also been increasing in popularity[3,4], with the bulk of these identifications being performed on pure compounds or mixtures of low variability (in terms of both the chemical and physical compositions of the sample). As a result of the reliance upon pattern-recognition procedures, however, both qualitative and quantitative analysis in the near-IR region can be complicated by the *false-sample problem*[2,5].

The false-sample problem arises whenever a pattern-recognition method is presented with a sample unlike any the method has ever analyzed before. In the regression-based analysis of near-IR spectral data, the false-sample problem arises when a sample must be analyzed whose composition is outside the domain of the samples used to develop the calibration equation. For example, the false-sample problem can appear as a result (1) of trace contamination of the test sample, (2) of gross substitution of one sample component (one that was not present in the training-set samples) for another component (one that was present in the training-set samples), (3) of instrumental drift or sampling difficulties, or (4) simply as the result of a component concentration either rising above or falling below the range of concentrations used in the training set. In quantitative near-IR spectrometry, when multiple linear regression is used to develop a prediction equation, any amount of

extraneous signal at the analytical wavelengths (regardless of its source) generates a corresponding change in the predicted analyte value. When complex mixtures of high variability are qualitatively identified using distance or direction metrics defined in an analytical-wavelength space, these extraneous signals at the analytical wavelengths again produce changes, this time in the distance or direction values used to make a sample identification. In other words, false samples can give rise both to erroneous analyte-concentration determinations and to sample misidentifications without any indication of the error.

Qualitative analytical methods designed for the detection of false samples using their near-infrared spectra have been proposed previously for quantitative near-IR spectrometry[6,7]. These detection methods allow different calibration equations to be employed automatically in the analysis of different kinds of samples (provided that training samples are available for each sample type). Even if the false sample cannot be identified as belonging to any known training set, an operator can still be alerted to the fact that the false sample cannot be predicted by the current calibration equation. "Bad" samples can thus be removed from a process by false-sample detection. Alternatively, a number of false samples can be identified, collected, and analyzed to produce a training set that describes the previously unknown sample type. To date, however, false-sample detection in qualitative near-IR spectrometry itself has been largely unexamined. This report describes an extension to the Bootstrap Error-Adjusted Single-sample Technique (BEAST)[7] that can be utilized for false-sample detection in both quantitative and qualitative near-IR spectrometry.

Basis of the Method. The basic BEAST is an experimental clustering technique for exploring multivariate data distributions. The technique considers each analytical wavelength to be a spatial dimension, so that spectra recorded at n wavelengths are represented as single points in an n-dimensional hyperspace. The magnitude of the signal observed at each wavelength is represented by the translation of the spectral point along each axis from the origin. Spectra of similar compounds produce clusters of points in similar regions of hyperspace as a consequence of this representation. A confidence-limit surface can be placed on a training-set spectral cluster by the BEAST (usually at three standard deviations, or 3 SDs, from the center of the training set), and test samples whose spectra project as points inside this surface are said to be of the same type as the training samples. Test-sample spectra that project outside the 3 SD surface ("false-sample" spectra) are not classified as members of the training-sample set.

To this point, the BEAST is conceptually similar to discriminant analysis using Mahalanobis distances[6]. There are, however, two principal differences between the methods:

1. The BEAST standard deviation (SD) can be symmetric or asymmetric. The Mahalanobis metric has often been referred to as a "rubber yardstick" whose length in hyperspace depends upon the orientation of the stick. The "stretch" of the Mahalanobis distance is symmetric (i.e., grasping the yardstick at both ends and pulling produces identical increases in the lengths of the upper and lower halves of the stick). The stretches of the upper and lower halves of the BEAST yardstick are not necessarily equal and depend, in fact, upon the

skew of the training set in the direction of the yardstick. This is an important capability to have in a metric when subclusters exist in the training set, when the response in one or more dimensions is nonlinear[5] , etc.

2. The BEAST is easily implemented on multiple-processor systems. The BEAST is a simple algorithm whose basic operations are random shuffling, sorting, and distance measurement. These simple operations are relatively easy to distribute among a number of processors[7].

Both differences between the BEAST and Mahalanobis metrics are the result of the nonparametric (i.e., making few assumptions about the nature of the underlying data distribution) nature of the BEAST. The BEAST comprises three operations:

1. A training set is carefully constructed from known samples (samples that have been analyzed or identified by some other reference procedure) in a way that adequately describes all possible sample variations. This step is common to most near-IR procedures.
2. A randomly selected set of samples (containing the same number of elements as the training set) is drawn from the training set, with replacement from the training set, to calculate the "bootstrap distribution"[8]. The Monte Carlo approximation to the bootstrap distribution is formed by a process involving repeated drawing of randomly selected sets from the training set. The bootstrap distributions of two mixtures are shown in Figure 1.

3. The bootstrap distribution from step 2 is used to estimate the population distribution for the training set. Each test-sample spectrum is projected into the same space as the bootstrap distribution, and a line is formed in hyperspace connecting the center of the bootstrap distribution and the test-sample spectral point. A hypercylinder formed about this line contains a number of points from the bootstrap distribution. The coordinates of the points within the hypercylinder are transformed into distances from the center of the bootstrap distribution, and these distances are projected onto the hyperline at the center of the hypercylinder. The projected distances form a univariate distribution whose quantiles are used to construct confidence limits in the direction of the hyperline.

Consider the following hypothetical process problem: a small plant, perhaps a pharmaceutical facility, receives raw materials in large drums and also removes its waste materials and garbage in somewhat similar drums. Naturally, the drums are color-coded and labeled. But suppose that one night a new employee is sweeping the plant and, having filled his dustpan, inadvertently empties the dustpan into the wrong container. The next day, this cubic meter of reagent is used to produce pharmaceutical capsules.

Particle-size variations, noise, drift, trace contamination, and other factors can all create a situation in which a simple discriminant test that places a single sample inside a training-set cluster does not necessarily indicate that this sample appropriately belongs to the training set. Instead, the sample can be a member of a new, slightly

different cluster that overlaps the training set. The hypothetical process problem could demonstrate this overlap effect: a single intact capsule, prepared from a reagent container contaminated with floor sweepings, would be likely to pass a near-IR examination based upon ordinary discriminant analysis.

Step 3 of the basic BEAST algorithm provides for the qualitative analysis of a single test sample. An alternative to step 3 utilizing multiple test samples provides a powerful extension to the BEAST algorithm capable of solving problems such as the hypothetical process problem proposed. In effect, this extension creates density contours inside of a training-set spectral cluster and detects perturbations of these contours using the bootstrap procedure (step 2). Thus, false samples can be detected as subclusters well inside the 3 SD limit of a training set. The accurate detection of subclusters makes possible "trace" near-IR analyses and analyses using a very small number of wavelengths.

THEORY

The Quantile BEAST is a flexible clustering procedure that can actually be implemented in a number of ways[7]. Extending the method to search for subclusters within a training set requires the formation of a training set **T** and a test set **X**, as well as the calculation of these sets' respective bootstrap distributions, **B** and $\mathbf{B}_{(\mathbf{X})}$. The discussion that follows outlines one route to a solution to the subcluster problem.

A training set of sample spectral values (e.g., reflectance, absorbance, etc.), recorded at d wavelengths from n training samples, is

represented by the n-by-d matrix T. (Generally, another n-by-d matrix V, containing validation samples, is also assembled from the same source as the training set. The sample set V serves as an indicator of how well the training set describes its overall population variation.) This is essentially the step 1 described earlier for the basic BEAST.

Step 2 of the basic BEAST calls for the calculation of bootstrap distributions. Bootstrap distributions can be calculated by an operation κ^7; $\kappa(T)$, $\kappa(X)$, and $\kappa(V)$ are each calculated in this manner. The results are the m-by-d arrays B, $B_{(X)}$, and $B_{(V)}$. The operation $\kappa(T)$, for example, begins by filling a matrix P with sample numbers to be used in bootstrap sample sets $B_{(*)}$:

$$P = p_{ij} = r \tag{1}$$

The values in P are scaled to the training-set size by:

$$P = [(n-1)P + 1] \tag{2}$$

A bootstrap sample $B_{(*)}$ is then created for each row i of the m-by-d bootstrap distribution B by

$$B_{(*)} = t_{Kj} \tag{3}$$

where K are the elements of the ith rows of P. The q-th row of B is filled by the center of the q-th bootstrap sample

$$b_{qj} = \sum_{i=1}^{n} b_{(s)ij} \; / \; n \tag{4}$$

and the center of the bootstrap distribution is

$$c_j = \sum_{i=1}^{m} b_{ij} \; / \; m \qquad (5)$$

The operation κ is then repeated using **X** and **V**.

The multivariate data in the bootstrap distributions are then reduced to a univariate form:

$$s_{(T)i} = \left(\sum_{j=1}^{d} (b_{ij} - c_j)^2 \right)^{1/2} \qquad (6)$$

$$s_{(X)i} = \left(\sum_{j=1}^{d} (b_{(X)ij} - c_j)^2 \right)^{1/2} \qquad (7)$$

$$s_{(V)i} = \left(\sum_{j=1}^{d} (b_{(V)ij} - c_j)^2 \right)^{1/2} \qquad (8)$$

and these distances are ordered and trimmed according to a trimming-index set

$$P_{(T)} = \{mp+1,\ mp+2,\ mp+3,\ \ldots,\ m-mp\} \qquad (9)$$

to reduce the leverage effects of isolated selections at the extremes of the bootstrap distributions. A hypercylinder can be formed about the line connecting **C** to the center of $B_{(X)}$ or $B_{(V)}$, giving directional selectivity to the information in $S_{(T)}$, $S_{(X)}$, and $S_{(V)}$ if desired[7]. $S_{(T)}$, $S_{(X)}$, and $S_{(V)}$ then have n_h elements instead of m elements. Subclusters can be detected without this selectivity, however. While this directional selectivity adds to the sensitivity of the subcluster

test, it also introduces a number of additional questions, e.g.: How small a radius is too small? How many replicates are required for a given radius? At what point is the additional sensitivity merely reacting to the particular training set selected, and not to any population characteristic? (Answers to some of these questions are examined below, and some others are discussed in reference 7.)

Cumulative distribution functions (CDFs) for quantile-quantile plotting (see Figure 2) are then formed by:

$$C_{(t)} = \partial(S_{(T)}P_{(T)}, S_{(T)}P_{(T)}) \tag{10}$$

$$C_{(X)} = \partial(S_{(T)}P_{(T)}, S_{(X)}P_{(T)}) \tag{11}$$

$$C_{(V)} = \partial(S_{(T)}P_{(T)}, S_{(X)}P_{(T)}) \tag{12}$$

Plotting the elements of $C_{(T)}$ on the abscissa versus the elements of either $C_{(X)}$ or $C_{(V)}$ on the ordinate produces a standard quantile-quantile (QQ) plot. Patterns in such a plot can be used to analyze structure in the spectral data[9], and the significance of the correlation between $C_{(T)}$ and $C_{(X)}$ can be used as an indication of the existence of subclusters in the spectral data. In the QQ plot, a straight line with unit slope and an intercept of 0 indicates that the two cumulative distribution functions are essentially identical (this should be observed when $C_{(V)}$ is on the ordinate). In the extended BEAST QQ plot, the presence of breaks in the line indicates that the CDF on the ordinate is multimodal (i.e., that the test set and training set of samples are not the same) (see Figure 3). Sharp bends in the QQ line also indicate the presence of more than one distribution.

EXPERIMENTAL

Equipment. Spectral data for all of the experiments were collected using a Technicon InfraAlyzer 400 filter spectrophotometer. The spectrophotometer was directly connected to a VAX 11/780 computer (Digital Equipment Corp.), and the spectral data were recorded at 18 wavelengths. The extended BEAST algorithm was implemented in Speakeasy IV Delta (VMS version, Speakeasy Computing Corp., Chicago, IL). The intact pharmaceutical capsules used as samples were individually scanned using a 90° conical aluminum holder/reflector designed for use with the spectrophotometer[5].

Materials. Samples were prepared from Maximum Strength Anacin-3 capsules (500 mg acetaminophen, Whitehall Laboratories, Inc., New York). These capsules have a blue end (cap) and a white end (body). In the contamination experiments the training samples were unadulterated Anacin-3 capsules; however, the capsules were emptied and repacked so that sample variations introduced by the repacking process appeared uniformly in the training, validation, and test sample sets. The validation samples were likewise unadulterated, repacked Anacin-3 capsules. The training sets each contained 10-13 capsules, depending upon the experiment involved, and the validation sets contained an equal number of samples.

Adulterants in the test-sample capsules were selected to simulate certain process-control problems. The first adulterant was aluminum dust (finest powder, Fisher Scientific Co., Fairlawn, NJ). The aluminum powder was blown into empty capsules and the amount that adhered to the capsule walls was determined by weighing each capsule before and after

the addition of the dust. The second adulterant was ordinary dust: floor sweepings were obtained from the dry bag of a wet/dry vacuum cleaner. A sizeable amount of this dust consisted of fibers (probably both natural and synthetic). This material was introduced into the capsules by emptying a number of test capsules into a beaker, mixing the material into the powder in the beaker, and repacking the capsules. By weighing the total masses of the analgesic powder and the floor sweepings, the average mass of floor sweepings per capsule was determined.

Types of Experiments Performed. The first experiment involved a training set composed of 13 intact Anacin-3 capsules. The first step was the determination of the adequacy of this training set by using a validation set. Bootstrap samples[7,8] were drawn from a 13-sample validation set and analyzed with the extended BEAST procedure described above. The standard deviation of the product-moment correlation coefficients calculated from the extended BEAST applied to these 26 samples was 0.01. The mean correlation coefficient was 0.99.

Once a validated training set was available, the first experiment sought to determine the effects of test sets with different locations and scales on the QQ plot and to ascertain the correlation coefficient returned by the extended BEAST. Preparing contaminated capsules sets that project at precisely the location and scale in hyperspace desired by an experimenter is nearly impossible, so the test-sample sets were created by the computer. The location and scale of the Anacin-3 training set were determined. Normally distributed pseudorandom numbers were then generated by the computer and these numbers were relocated and scaled to have the desired relationship to the Anacin-3 training set. In every case, the test set and the training set each contained 13 samples. Three

types of relationships between the training and test sets were explored in this manner:

1. The effects of pure location differences between the training set and the test set were investigated. A test set was created of the same size as the training set in spectral hyperspace. Initially the test set and the training set also shared the same center in hyperspace. As the sets were moved apart, the effects on the QQ plot and the correlation coefficient calculated from the plot were monitored.

2. The effects of pure scale differences between the training set and the test set were examined by creating test sets with different sizes in hyperspace. The test set and the training set shared the same center in hyperspace in these program runs. Two different types of behavior occurred in the QQ plots in this case. The first type of behavior was observed when the training set was larger than the test set, and the second type was observed when the training set was smaller than the test set. Again, the patterns in the QQ plots and the correlation coefficient were the indicators of the different types of behavior.

3. The effects of simultaneous differences in location and scale on QQ plots and correlation coefficients were determined. Two types of behavior were again observed, one when the training set was larger than the test set, and the other when the training set was smaller.

The second experiment determined the bias and RSD for the subcluster-detection procedure as a function of the number of training samples used, the number of bootstrap replications employed, the number of wavelengths monitored (the dimensionality of the hyperspace), and the radius of the hypercylinder. This experiment used computer-generated data for both the training set the and test set. The training set and the test set each had the same number of samples in a given program run. New normally distributed pseudorandom numbers were generated for both the training set and the test set in each program run.

The third experiment looked at real Anacin-3 capsules with simulated process-control problems. The first problem was the contamination of capsules with aluminum dust. The test set and the training set each contained 10 Anacin-3 capsules. The test-set capsules contained an average of 208 μg of aluminum dust per capsule (σ=136 μg), and the average total capsule mass was 704 mg (σ=32 mg). The second simulated process-control problem was described above -- the detection of floor sweepings in capsules. A beaker containing 6.1052 g of analgesic powder from Anacin-3 capsules was contaminated with 0.00135 g of floor sweepings. Ten capsules were then packed with this material, giving an average concentration of 221 ppm floor sweepings in the capsules.

RESULTS AND DISCUSSION

The results of the first experiment describe the behavior of the subcluster-detection procedure as the location and scale of a test set of samples varies with respect to a fixed training set. Each of the location and scale tests used the same fixed training set of 13 Anacin-3

capsules and a different computer-generated set of test samples. The construction of the pure location difference test is depicted in Figure 4. The test set and the training set are each depicted as a rectangle in Figure 4 (the actual sets were not rectangular, however; the rectangular shape was chosen arbitrarily to emphasize that the actual shape of the spectral data sets in hyperspace is not important to the subcluster-detection procedure). The training set and test set are exactly the same size, and differ only in the location of their centers. The training set is represented by the rectangle on the left, while the test set is represented by the overlapping rectangle on the right. In the actual experiment, the centers of the training and test sets were 0, 0.25, 0.5, 1, 1.25, 1.875, 2.5, 5, and 10 SDs apart. (The distance in SDs refers to the SD of either the training set or the test set. The fact that the two sets are the of the same scale and the fact that each set was drawn from a population that was spherical in shape means that the SD of either set can be used in measuring the distance between the two set centers.)

The effect on the QQ plot of a pure location difference between the test set and the training set is shown in Figure 5. The presence of a distinct subcluster in the training set causes a break in the line of the QQ plot. The test set forms the lower line and the training set forms the upper line. (The ordinate contains data from the test set and the training set, while the abscissa contains only data from the training set. Thus, the axis range of the abscissa gives the range of the training set, which can then be compared to the values along the ordinate to determine which group is the training set and which group is the test set.) In Figure 5 the centers of the test-set and training-set distributions are 2.5 SDs apart. As the distance between the two centers

is increased beyond 2.5 SDs, the slopes of the upper and lower lines are reduced and the gap between their ends increases. As the distance between the two centers is reduced below 2.5 SDs, the slopes of both of the lines increases and the gap between the ends of the lines shrinks. When the distance between the two centers is 0.5 SD or less, no gap between the lines is distinguishable, and the test set behaves much like a validation set.

The effect of a pure location difference between the test set and the training set on the product-moment correlation coefficient calculated from the distributions in the QQ plot is shown in Figure 6. The dotted line in the figure represents a lower confidence limit (2 SDs below the mean correlation coefficient, or the 98% level) calculated on the training set using validation samples. Above the dotted line the training set and the test set are indistinguishable, and the test set can be said to be composed of the same type of samples as the training set. When the correlation coefficient is less than approximately 0.96, the test set can be said to be different from the training set at the 98% level. The most important thing to note is that when two otherwise equivalent groups are separated by more than approximately 0.8 SD, the subcluster-detection procedure is able to signal the presence of a false-sample set.

The next type of effect that can occur is caused by pure scale differences between the test set and the training set. The scale-difference effect on the subcluster-detection procedure depends upon whether the training set is larger than the test set, or the test set is larger than the training set. In both cases the scale-difference effect was investigated by creating test sets that shared the same center in

hyperspace with the training set of Anacin-3 capsules. The scale of the computer-generated test set was then expanded or contracted, and the effect on the QQ plot and correlation coefficient produced by the subcluster-detection procedure was analyzed.

The first case of the pure scale-difference effect (training set larger than test set) is depicted in Figure 7. Figure 8 shows a QQ plot generated with a test set that is a factor of 5 smaller than the training set. The presence of a distinct subcluster in the spectral data on the ordinate causes the two sharp bends in the plot: the center line in the QQ plot is generated by the test set, which separates the two end segments that correspond to the training set. As the test set becomes smaller, the slope of the center line segment decreases (although when the test set is beyond a factor of 5-10 times smaller, the approach to a slope of zero slows considerably). As the test set grows larger, the slope of the center line increases and the two "bends" begin to disappear. By the time the test set is only 1.25 times smaller than the training set, the slope of the center line has increased to the point where it is difficult to tell if the center line segment is real.

Figure 9 describes the effect of the scale difference (when the training set is larger than the test set) on the correlation coefficient calculated for the QQ plot. The dotted line again represents the 98% level calculated on the training set using validation samples. When the test set is more than a factor of approximately 1.5 times smaller than the training set, the test set is flagged as a false-sample set even though the two sets share exactly the same center in hyperspace.

The inverse pure scale-difference situation occurs when the test set is larger than the training set in scale (see Figure 10). In this case,

the same sort of sharp bend actually appears in the QQ plot when a false-sample situation is present, except that now the line segment with the small slope corresponds to the training set, and the line segment appears at one end of the plot instead of in the center. In Figure 11, the position of the training set at the end of the plot reduces the sensitivity of the QQ plot and the correlation coefficient to changes in test-set scale. In Figure 11 the test set is a factor of 2.5 larger in scale than the training set. The upper line corresponds to the training set and the lower line to the test set. As the test set shrinks further in scale the slope of its line segment decreases. When the difference in scales is less than a factor of approximately 2.25 and the test set and training set share the same center, the test set appears to belong to the training-set population. As the test set scale is increased beyond the factor of 2.5, the upper line (training set) appears to shrink in both slope and length, eventually approaching the top edge of the plot.

The simultaneous shrinkage of the slope and the length of the training-set line segment as the test set grows larger tends to preserve an overall linearity in the QQ plot. In other words, a sharp bend in the plot means more to the correlation coefficient when it occurs at the center of the plot than at the end of the line. The change in the correlation coefficient as a function of the test-set scale (see Figure 12) reflects this property of the correlation coefficient. The curve in Figure 12 is the most gradual of all of the curves discussed to this point, crossing the 98% level for false-sample set detection only when the test set is more than a factor of approximately 2.3 times larger than the training set.

In order to get useful information from a single correlation coefficient calculated in this subcluster-detection procedure, one must know whether the training set or the test set is larger in scale. A cursory examination of the axis scales in the QQ plots presented thus far reveals one simple approach to the problem of determining the relationship between the set scales. The relationship between the test-set and the training-set scales can be determined from the coefficients of the best-fit straight line through the QQ plot. (Linear fitting is commonly performed with QQ plots because the slope of the line contains information about the variance or spread of the data, and the intercept of the line contains information about the mean of the data[7,8].) In a subcluster-detection situation (i.e., spectral data appearing near or below the 3 SD limit of a training set cluster), when the test set is smaller than the training set the slope of the best-fit line through the QQ plot will be small (between zero and one). When the test set is larger than the training set, larger slopes (greater than one) will be observed. It should be noted that large distances (more than 3 SDs) between the test-set and training-set centers will also produce large slopes in the best-fit line to the QQ plot. However, these distances do not represent false samples that would appear as training-set subclusters, and such samples would be analyzed using the conventional BEAST metric[7].

In most real applications of the subcluster-detection method, a simultaneous difference in both the location and the scale of the test set would be expected in a false-sample situation. Once again, two types of behavior are observed in the method, one when the test set is larger

than the training set, and another when the test set is smaller than the training set.

Figure 13 represents schematically the case in which the test set is larger than the training set, and the difference between the two set centers varies. Figure 14 shows a QQ plot with a computer-generated test set that is only a factor of two larger in scale than the Anacin-3 training set. The center of the test set and the training set are only 0.5 SD apart (measured using the smaller SD of the training set). The test set forms the upper line and the training set forms the lower line. Despite the small differences in the scale and location of the two groups, a clear break in the QQ plot is evident, indicating that a false-sample situation exists. As the test set becomes larger the slope of the upper line increases. As the distance between the test set and training set increases the size of the gap between the two groups also increases. The location-difference effect and the scale-difference effect act in concert to make the subcluster-detection method sensitive to very small differences between the spectral groups. Even when the two groups differ in size by a factor as small as 2, the QQ plot suggests the presence of a subcluster in the data on the ordinate regardless of the difference in the locations of the two group's centers.

Figure 15 describes the effect of location changes on the correlation coefficient from the QQ plot when the test set is larger than the training set. Obviously, the more alike the test set and training set are in terms of scale the larger the correlation is between them. A simple statistic such as the correlation coefficient is adequate for describing the behavior of the QQ plot when the test set is larger than the training set: the curves in Figure 15 are fairly uncomplicated, and

little confusion over whether a test set is a false-sample set seems likely to arise (indeed, the only time that the test set fails the 98% level training-set population test occurs when the distance between the two set centers is less than 0.2 SD).

When the test set is smaller than the training set and the distance between the two set centers varies (see Figure 16) the QQ plot is still reasonably easy to interpret. The QQ plot has an inverse sigmoidal shape composed of three lines and appears similar to the plot in Figure 8, where the test set is also smaller than the training set. The line with the lowest slope is still the test set, but this line does not occur in the center of the QQ plot. Instead, the center of the test set line "slides" with constant slope along the imaginary y=x line that would exist if the test set and the training set were drawn from the same population. The combined effects of the location change and the scale difference between the two sets once again makes the QQ plot quite sensitive to minor differences that might indicate a false-sample situation. When the test set is smaller than the training set by a factor of 5 or 10, the false-sample nature of the test set is apparent in the QQ plot regardless of the distance separating the two set centers. When the test set is smaller than the training set by a factor as small as 2, the false-sample nature of the test set is still apparent until the distance separating the set centers is only 0.1 SD.

Analyzing the behavior of the correlation coefficient when the test set is smaller than the training set and when the centers differ in location is more complicated than explaining the QQ plot appearance (see Figure 17). Two peaks appear in the graph of the correlation coefficient when the test set is smaller than the training set and the centers of the

sets are slowly drawn apart. The first peak reflects a breakdown in the symmetry of the spectral sets in hyperspace and consequently in the QQ plot. At this point the correlation coefficient enters a "bend regime" where the difference between set scales dominates the correlation coefficient calculated from the QQ plot. As discussed earlier in connection with Figures 11 and 12, the "bends" introduced by different set scales have a larger effect on the correlation when the bends occur in the middle of the line than when the bends occur at the end. Thus, after an initial drop in the correlation coefficient caused by its entering the bend regime, the correlation coefficient begins to rise again as the test set and training set continue to move apart. However, eventually the difference in the locations of the set centers must begin to exert an effect on the correlation coefficient. At this point the curves enter the "break regime", forming a second peak in the graph. In the break regime, the difference between the centers of the training set and test set dominates the correlation coefficient. Larger differences between the scale of the training set and test set cause later transitions into the break regime. In Figure 17 the beginning of the bend regime is evident in each curve when the distance between the training set and test set centers has reached 0.1 SD. When the test set is a factor of 2 smaller than the training set, the break regime begins when the distance between the two set centers is around 1 SD. When the test set is a factor of 5 smaller than the training set, the break regime does not begin until the distance between the sets has reached 2 SDs. When the test set is a factor of 10 smaller than the training set, the start of the break regime does not begin until the distance between the centers of the sets is beyond the 3 SD training-set limit. Despite the

effects of the regime shifts, the subcluster-detection method accurately flags as false samples test sets whose scales differ from the training set by a factor of 5 or 10, regardless of their distance from the center of the training set.

When the test set is larger than the training set (probably the more common experimental situation), the break regime occurs very early and dominates the curves (see Figure 15), and the regime-transition peaks overlap. Only when the training set is a factor of 10 smaller than the test set is the break regime transition peak delayed long enough to be resolved in Figure 15.

Bias and RSD of the Subcluster-Detection Procedure. In order to use the correlation coefficient from the QQ plot to describe the relationship between a training set and a test set, one should be aware of how the correlation coefficient changes with commonly adjustable experimental parameters. The number of training samples used (and test samples, as these are coupled in the extended BEAST subcluster-detection method to simplify the testing procedure), the number of bootstrap replications employed, the number of wavelengths monitored, and the radius of the hypercylinder all affect the bias and RSD of the correlation coefficient. Using computer-generated data for a training set and a test set that were both drawn from the same population (a nonfalse-sample situation that should produce a correlation coefficient of 1), bias and RSD plots for the four common experimental variables were created.

Figure 18 shows the bias and RSD of the correlation coefficient as a function of the number of training (and test) set samples. Each point comprises six trials using two wavelengths and all 1000 bootstrap replications (no hypercylinder was used to screen out replicates in

certain directions). Both the bias and the RSD decrease rapidly with an increasing number of samples, gradually reaching a plateau around 0.2-0.3% by the time 100 samples are used. This decrease is not in itself particularly surprising, especially when considered in the light of previous results with the single-sample technique[7]. What is somewhat surprising is that the largest bias and error values, obtained using only 10 samples, are still less than 3%.

The effect of the number of bootstrap replications on the correlation coefficient is described in Figure 19. Each point represents the mean of six trials using two wavelengths and 100 training samples. Using bootstrap replications in QQ plots to detect subclusters amounts to smoothing the training and test sets before plotting. At first glance it seems impossible to oversmooth a spectral data set in hyperspace using a bootstrap procedure. However, the bootstrap process involves drawing samples randomly and with replacement from a set to form a bootstrap set that is used to calculate a replicate point in the bootstrap distribution. Thus, as more bootstrap replicates are created, the likelihood of creating an extreme value increases. The increased likelihood of creating an extreme value may be responsible for the small increase in the bias and RSD observed at the largest number of bootstrap replications. However, the trimming index was introduced in equations 9-12 precisely to eliminate the extreme-value effect. All of these bias and RSD values are very small, and the slight rises in the bias and RSD at 1000 replications are probably just an artifact of the particular sample sets generated by the computer for those six trials.

In Figure 20 the bias and RSD of the correlation coefficient is given as a function of the number of wavelengths (or more properly the

dimensionality of the space) used in the analysis. Each point represents the mean of six trials using 100 training samples and 1000 bootstrap replications. Once again, the bias and RSD are quite small. The slow rise in the bias and RSD of the subcluster-detection method with increasing dimensionality is similar to the rise noted for the RSD of the single-sample technique[7]. As the dimensionality of the analytical space increases, the space becomes sparsely populated with points unless the number of points is also increased. In other words, to describe more variables, one needs more points.

Figure 21 depicts the relationship between the bias and RSD of the correlation coefficient and the radius of a hypercylinder formed about a line connecting the center of the training-set replicates to the center of the test-set replicates. Each point represents the mean of six trials using 100 training samples and 1000 bootstrap replicates in a two-dimensional space. The original idea behind the BEAST was to connect the mean point of a training set in spectral hyperspace to a test spectral point with a hyperline. The probability density along this hyperline would then be used to determine the probability that the test spectral point was actually a member of the population from which the training set was drawn. However, the probability-density function in hyperspace is difficult to calculate, so it is approximated by the bootstrap distribution created using the training set. Unfortunately, the bootstrap distribution is described by a finite number of points, none of which would be likely to fall on a hyperline connecting the centers of the test set and training set. The hypercylinder is an approximation to the hyperline that is required to define the density of the bootstrap distribution in the direction of the test-sample spectral point. The

approximation works well in the single-sample case, where the test statistic is the standard error in a specific direction. When the test statistic is the correlation coefficient between two groups, however, using a hypercylinder to exclude certain members of the groups seems to make the method excessively sensitive to fortuitous agglomerations of points. Certainly in Figure 21, the bias and RSD do rise as the hypercylinder radius is increased further from the theoretical hyperline with a radius of zero. By the time the radius is increased enough to include all of the replicate points, though, both the bias and the RSD have dropped to approximately 0.2% (see Figures 18-20).

Pharmaceutical Capsules with Simulated Process-Control Problems. A training set of ten repacked Anacin-3 capsules and a validation set of ten repacked Anacin-3 capsules were analyzed with the subcluster-detection method. The QQ plot line showed no breaks and only two slight bends near the ends of the line. The correlation coefficient for the line was 0.995, and the 98% level calculated using 20 bootstrap sets derived from validation samples was 0.966. When aluminum dust was added to ten repacked Anacin-3 capsules as described in the experimental section, the subcluster-detection method produced the QQ plot shown in Figure 22. This plot has a correlation coefficient of 0.795. The clear break in the line indicates that the aluminum-containing test-sample set was not drawn from the same population as the training-set samples. The average concentration of aluminum in the intact capsules was 296 ppm.

The hypothetical process-control problem discussed earlier, in which floor sweepings were accidentally introduced into a batch of pharmaceutical capsules, was also investigated with the new subcluster-detection method. Ten capsules containing an average of 221 ppm floor

sweepings per capsule were used to produce the QQ plot in Figure 23. The bend in the line suggests that the test set is really a false-sample set with respect to the training set. The upper line segment represents the test set, which is slightly larger in scale than the training set (as indicated by slope of the line segment itself and by the slope of the best-fit line through the entire QQ plot, which was 2.5). The test set is larger than the training set because the floor sweepings are not evenly distributed throughout the 10-capsule set (homogeneity is difficult to achieve with floor sweepings and acetaminophen powder). The correlation coefficient calculated for this plot was 0.963, which makes the sample set a false-sample set at the 98% level calculated above.

It is interesting to note that a simple t-test for the difference of the mean distance of the training samples from their center and the mean distance of the test samples from the training center does not perform as well as the subcluster-detection procedure using QQ plots. For the floor-sweepings data, a simple t-test predicts an 11% probability that the two mean set distances are the same. The t-test assumes that each set represents samples drawn from two normally distributed populations that share a common variance. When either assumption is violated, the t-test loses some of its power to differentiate between the training set and the test set. The subcluster-detection method using QQ plots makes no assumptions about the shapes or sizes of either set, and consequently retains its power to differentiate between the sets under a wider variety of circumstances.

CONCLUSIONS

False-sample spectra can be detected inside the 3 SD limit on a training-set spectral cluster when multiple test samples are available. The detection of small differences between a training set and a test set (subcluster-detection) is possible using the extended BEAST algorithm, even in some cases where the distance between the centers of the sets is zero. The subcluster-detection method provides additional sensitivity for analyses because overlap of data clusters is no longer a severe problem. Thus, near-IR analyses can be carried out successfully for constituents present in samples at low concentrations (a few hundred ppm), and with simple process sensors using only a small number of wavelengths (even 1 or 2). Perhaps most importantly, the subcluster-detection method provides a solution to the false-sample problem in both:

1. quantitative analysis, where it can assign a group of samples to a particular prediction equation if a training set corresponding to that group of samples available, and warn users if no such training set exists; and

2. qualitative analysis, where false-sample detection is necessary in cases such as library searching and "smart calibration" (where the computer selects a prediction equation by a library search). Factors like instrumental drift and trace contamination can generate errors that are too small to be noticeable in individual samples, yet are large enough to show up clearly in a sample set. The subcluster-

detection scheme employed in the extended BEAST is potentially applicable to both quality control and "smart calibration" problems. Investigations of these applications will be the subject of future reports.

CREDIT

This work was supported in part by the National Science Foundation through Grant CHE 87-22639, by the Office of Naval Research, by the Upjohn Company, and by Bran+Luebbe, Inc.

LITERATURE CITED

1. D. L. Wetzel, Anal. Chem. 55, 1165A (1983).

2. R. S. McDonald, Anal. Chem. 58, 1906 (1986).

3. H. Mark, Anal. Chem. 58, 379 (1986).

4. E. W. Ciurczak and T. A. Maldacker, Spectroscopy 1, 36 (1986).

5. R. A. Lodder, M. Selby, and G. M. Hieftje, Anal. Chem. 59, 1921 (1987).

6. H. L. Mark and D. Tunnell, Anal. Chem. 57, 1449 (1985).

7. R. A. Lodder and G. M. Hieftje, Appl. Spectrosc., "Quantile BEAST Attacks the False-Sample Problem in Near-Infrared Reflectance Analysis", submitted for publication, 1988.

8. B. Efron, Biometrika 68, 589 (1981).

9. E. Parzen, in *Some Recent Advances in Statistics* (Academic Press, London, 1982), pp. 23-52.

FIGURE CAPTIONS

Figure 1. Two thousand bootstrap-replicate spectra recorded at two wavelengths and projected as points in a two-dimensional space. Each point represents the center of a bootstrap sample set containing the same number of samples as the original training set. One thousand spectra form a training set, A, and one thousand spectra form a test set, B.

Figure 2. The formation of cumulative distribution functions (CDFs) from spectra similar to those in Figure 1. The abscissa values in Figure 2 represent the normalized Euclidean distances of each point in Figure 1 from the center of the training set A. Distributions like A and B that differ only in location form CDFs like F(1) and F(2). As p is varied between 0 and 1, a horizontal line intersecting F(1) and F(2) at level p selects pairs on the abscissa that can be used for quantile-quantile (QQ) plots.

Figure 3. A QQ plot of a bimodal distribution. When two distributions match in location, scale, and skew, the QQ-plot pattern is a straight line from corner to corner (i.e., a line with unit slope and zero intercept). The break in the line indicates the presence of two distributions with the same scale and skew, differing only in their locations.

Figure 4. Two distributions in spectral hyperspace, a training set and a test set. The two population distributions are identical except for their locations (centers). The shape of the distributions in this figure

have been arbitrarily selected to be rectangles. The test set rectangle and the training set rectangle overlap. In the actual experiment the training set consisted of 13 unadulterated Anacin-3 capsules, and the test set consisted of 13 samples that were computer-generated as a function of the Anacin-3 training set.

Figure 5. A QQ plot from the subcluster-detection method corresponding to the pure location-difference situation depicted in Figure 4. The ordinate contains data from the test set and the training set, while the abscissa contains only data from the training set. Thus, the test set forms the lower line segment and the training set forms the upper line segment in the QQ plot.

Figure 6. The effect of a pure location difference on the correlation coefficient calculated from a QQ plot. The abscissa values are distances in SDs of the training set. The dotted line represents a 98% limit on the training set calculated using validation samples. Correlations above the dotted line represent a match between the training set and test set population distributions. Correlations below the line represent a false-sample situation.

Figure 7. A training set and a test set in spectral hyperspace. The two population distributions share the same center, and the training-set population distribution is larger in scale than the test-set distribution. The distribution shapes depicted in the figure are arbitrarily selected.

Figure 8. A QQ plot from the subcluster-detection method corresponding to the pure scale-difference situation described by Figure 7. The test set is smaller in scale than the training set, and the test set forms the center line segment with the small slope.

Figure 9. The effect of a pure scale difference (test set smaller than the training set) on the correlation coefficient calculated from a QQ plot. The abscissa values represent the factor by which the test set is smaller in scale than the training set. The dotted line represents a 98% limit on the training set.

Figure 10. A training set and a test set in spectral hyperspace. The two population distributions share the same center, and the training-set population distribution is smaller in scale than the test-set population distribution. The distribution shapes shown in the figure are arbitrarily selected.

Figure 11. A QQ plot from the subcluster-detection method corresponding to the pure scale-difference situation described by Figure 10. The test set is larger in scale than the training set, and the test set forms the lower line with the larger slope in the figure. The bend in the line is slight because the difference between the set scales is only a factor of 2.

Figure 12. The effect of a pure scale difference (training set smaller than the test set) on the correlation coefficient calculated from a QQ plot. The abscissa values represent the factor by which the test set is

larger in scale than the training set. The dotted line represents a 98% limit on the training set.

Figure 13. A training set and a test set exhibiting simultaneous location and scale differences. The test-set population distribution is larger in scale than the training-set population distribution.

Figure 14. A QQ plot from the subcluster-detection method corresponding to the simultaneous location and scale differences situation represented in Figure 13. The test set is a factor of 2 larger in scale than the training set, and the two set centers are 0.5 SD (training set SDs) apart. The training set forms the lower line in the figure, and the test set forms the upper line.

Figure 15. The effect of simultaneous location and scale differences on the correlation coefficient calculated from a QQ plot when the test set is larger in scale than the training set. The abscissa values represent the distance between the two sets in terms of SDs of the training set. The horisontal long-dashed line represents a 98% limit on the training set. The short-dashed line represents a test set that is a factor of 2 larger than the training set, the dotted line a test set that is a factor of 5 larger than the training set, and the solid line a test set that is a factor of 10 larger than the training set.

Figure 16. A training set and a test set exhibiting simultaneous location and scale differences. The training-set population distribution is larger in scale than the test-set population distribution.

Figure 17. The effect of simultaneous location and scale differences on the correlation coefficient calculated from a QQ plot when the training set is larger than the test set in scale. The abscissa values represent the distance between the two sets in SDs of the training set. The horisontal long-dashed line represents a 98% limit on the training set. The short-dashed line represents a training set that is a factor of 2 larger than the test set, the dotted line a training set that is a factor of 5 larger than the test set, and the solid line a training set that is a factor of 10 larger than the test set.

Figure 18. The bias and RSD of the subcluster-detection method's correlation coefficient as a function of the number of training (and test) samples (using 2 wavelengths and 1000 bootstrap replications). The dotted line represents the bias and the solid line represents the RSD.

Figure 19. The bias and RSD of the correlation coefficient from the QQ plot of the subcluster-detection method as a function of the number of bootstrap replications employed by the method (using 100 training and 100 test samples, and 2 wavelengths). The dotted line represents the bias and the solid line represents the RSD.

Figure 20. The bias and RSD of the correlation coefficient from the subcluster-detection method as a function of the dimensionality of the spectral space (the number of wavelengths used). These data were created using 100 training samples, 100 test samples, and 1000 bootstrap

replications of each set. The dotted line represents the bias and the solid line represents the RSD.

Figure 21. The bias and RSD of the correlation coefficient from the subcluster-detection method as a function of the radius of the hypercylinder (an important single-sample method variable). These data were created using 100 training samples, 100 test samples, 1000 bootstrap replications of the training and test sets, and two wavelengths. The dotted line represents the bias and the solid line represents the RSD.

Figure 22. QQ plot from the subcluster-detection method created by Anacin-3 capsules containing an average of 296 ppm aluminum dust. The lower line corresponds to the training set and the upper line to the test set. Spectra were recorded at 18 wavelengths, but only one wavelength was needed in this analysis.

Figure 23. QQ plot from the subcluster-detection method created by Anacin-3 capsules containing an average of 221 ppm floor sweepings. The upper line segment with the larger slope corresponds to the test set, while the lower line segment corresponds to the training set. The spectra were analyzed using 18 wavelengths.

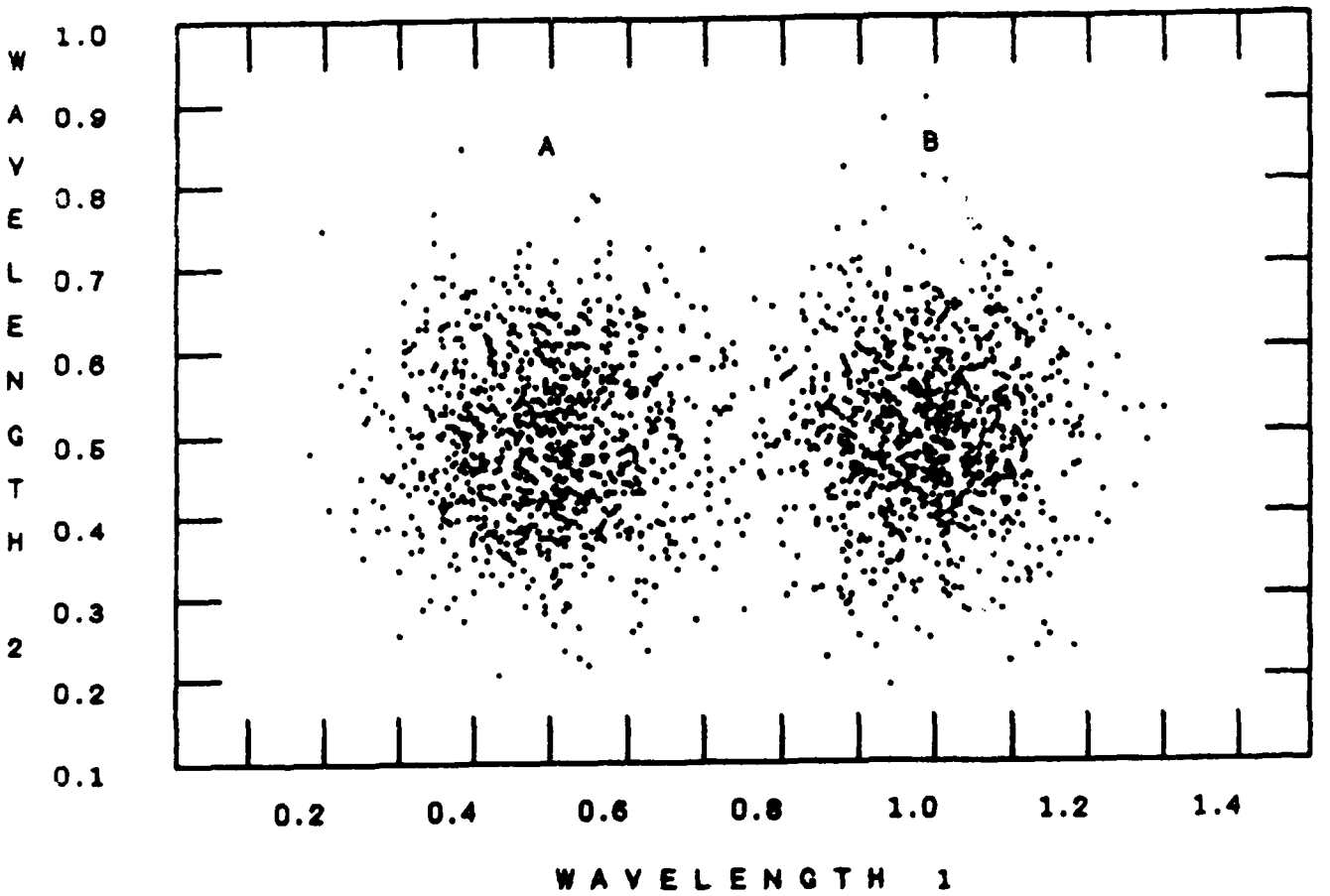
WAVELENGTH 2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
A
B
0.2
0.4
0.6
0.8
1.0
1.2
1.4
WAVELENGTH 1

Fig. 1

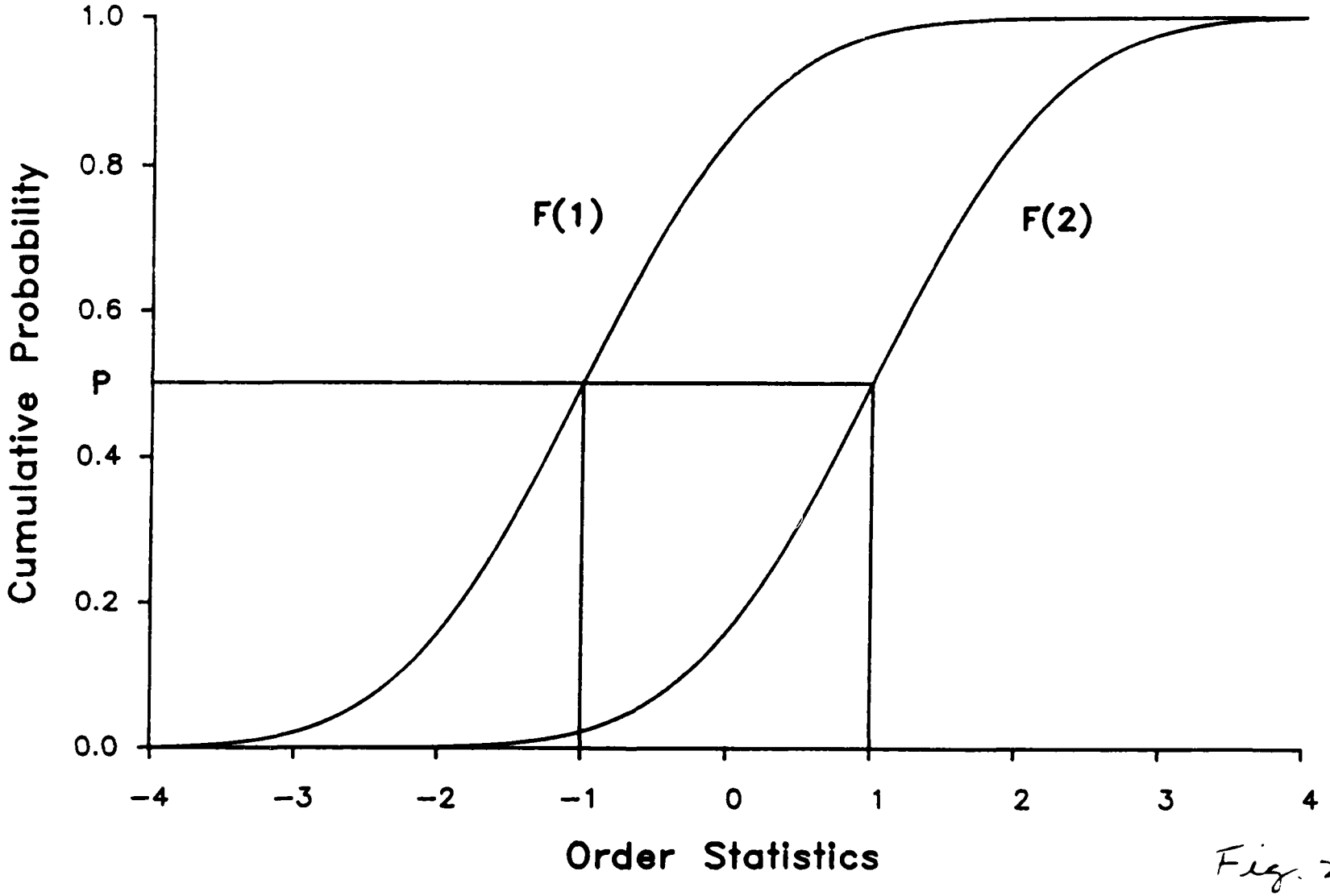
1.0
0.8
0.6
P
0.4
0.2
0.0
Cumulative Probability
F(1)
F(2)
−4
−3
−2
−1
0
1
2
3
4
Order Statistics

Fig. 2

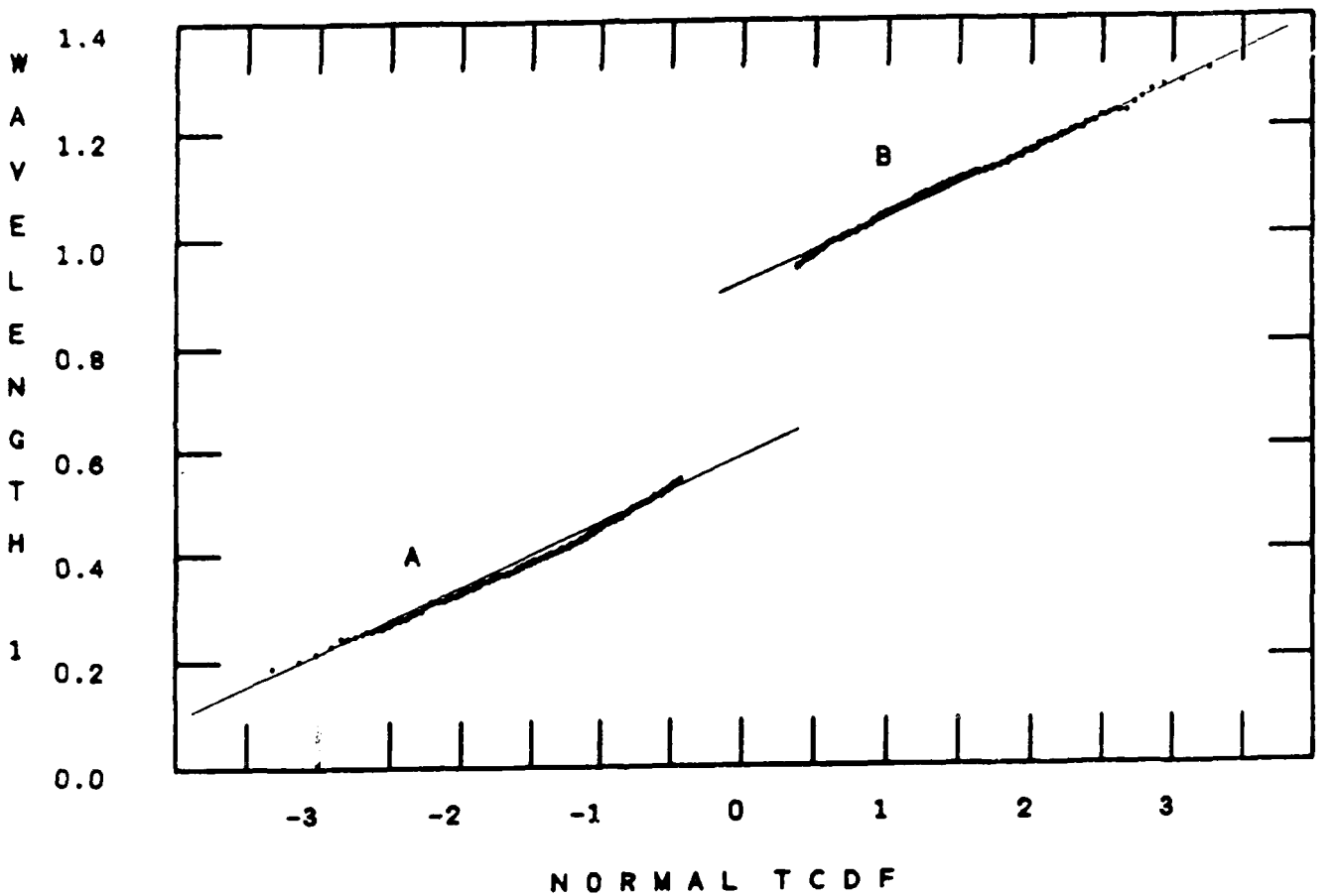

Fig. 3

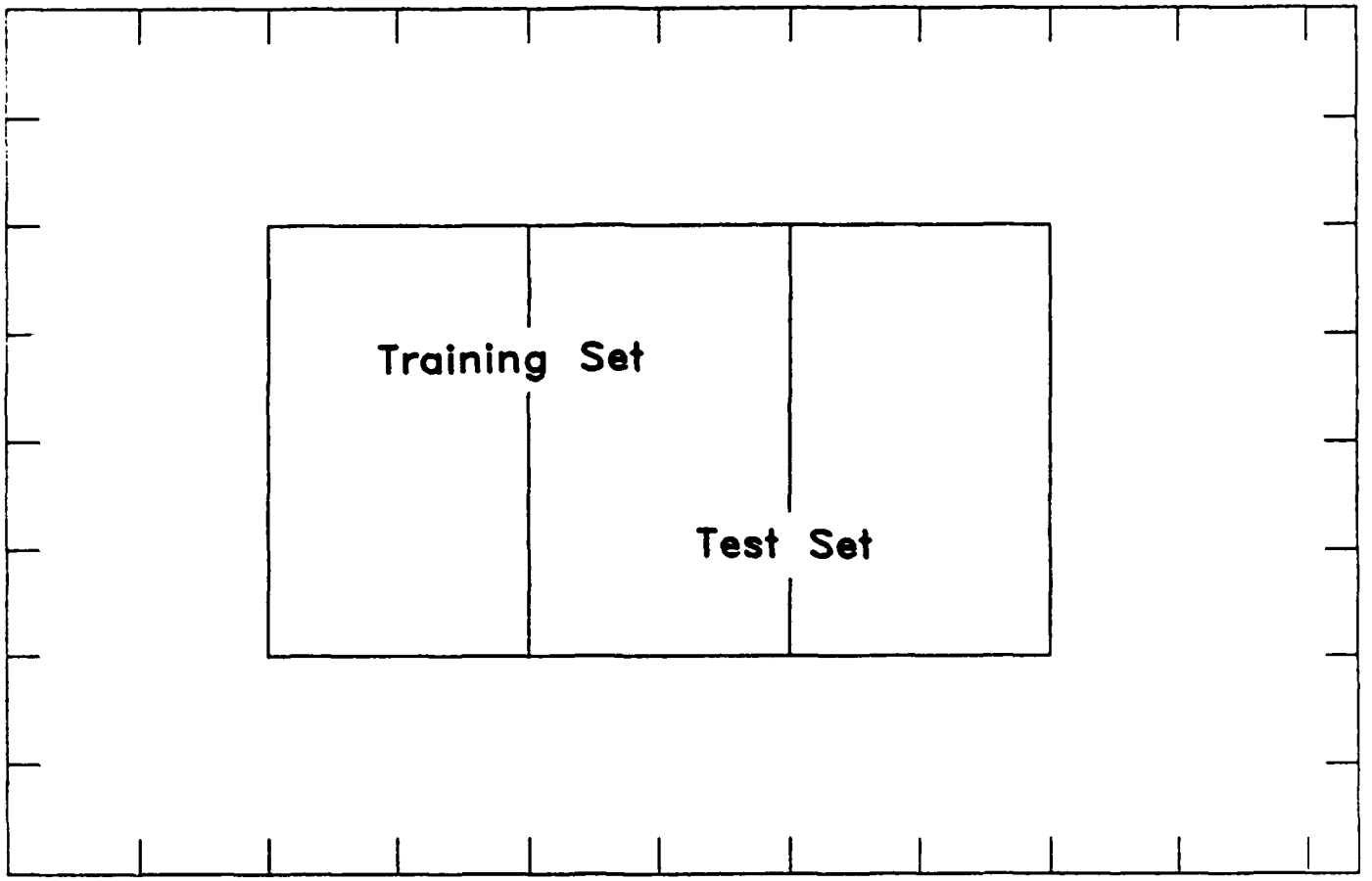

Fig. 4

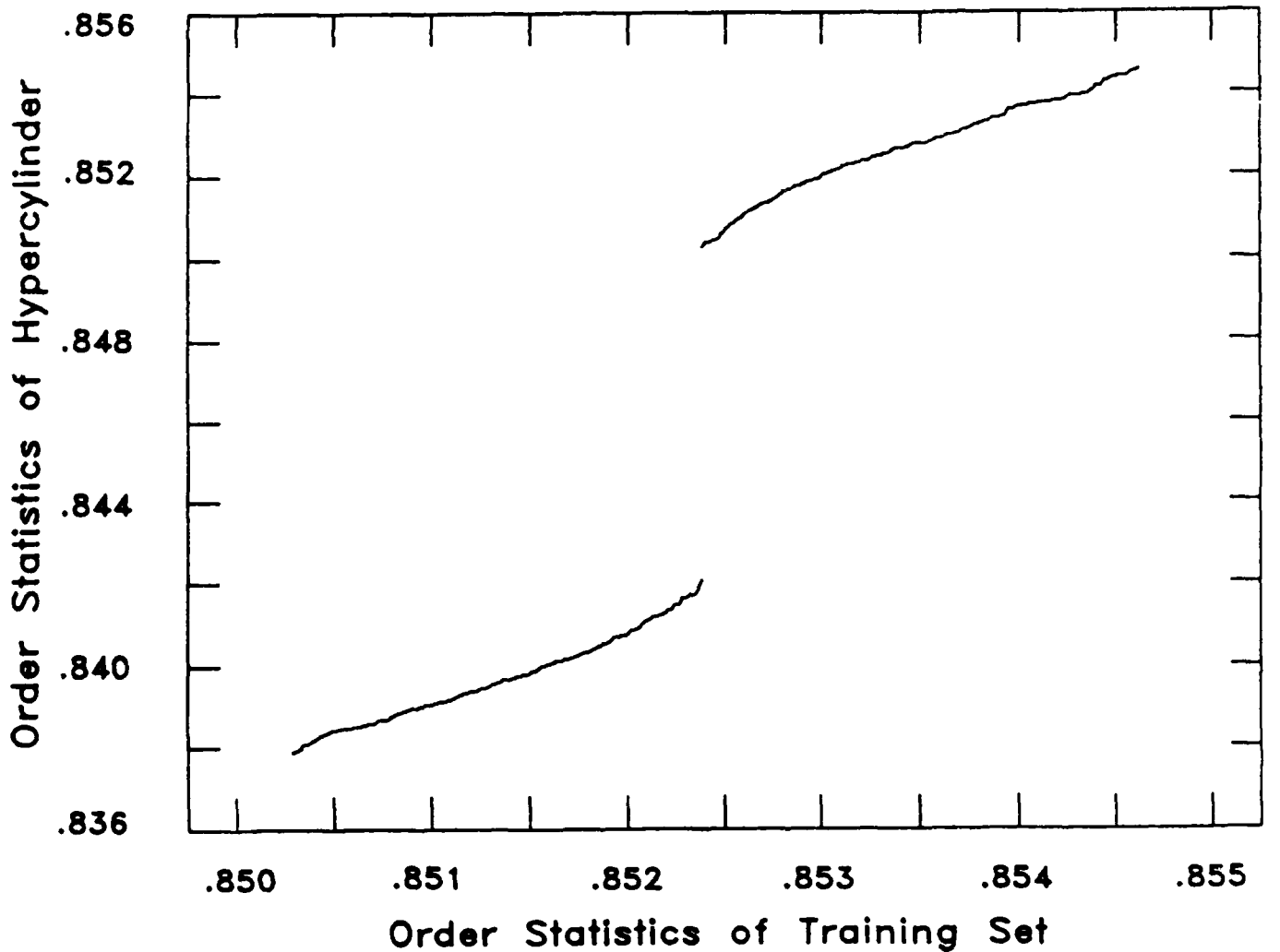
.856
.852
.848
.844
.840
.836
Order Statistics of Hypercylinder
.850
.851
.852
.853
.854
.855
Order Statistics of Training Set

Fig 5

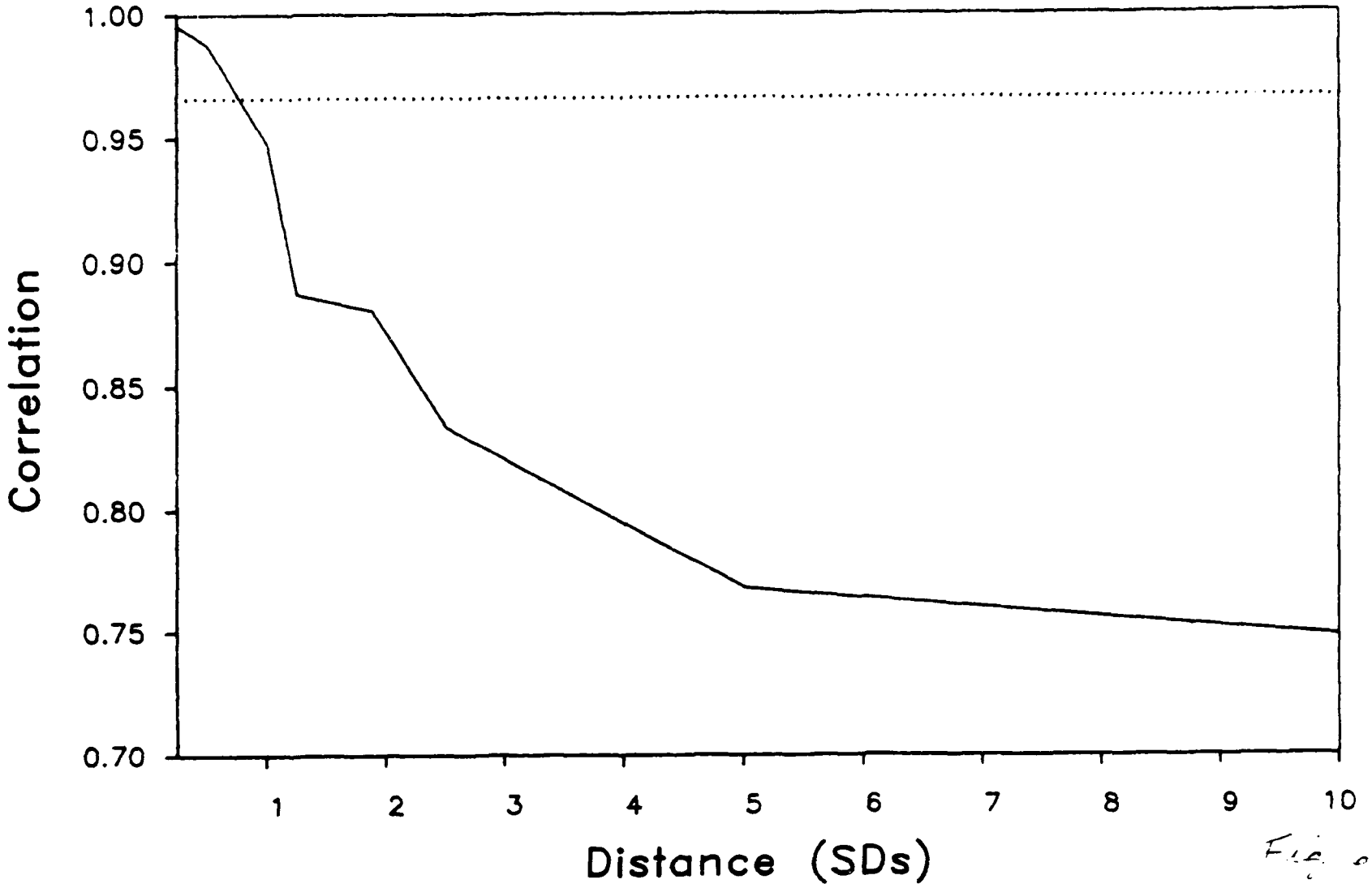
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Correlation
1
2
3
4
5
6
7
8
9
10
Distance (SDs)

Fig. 2

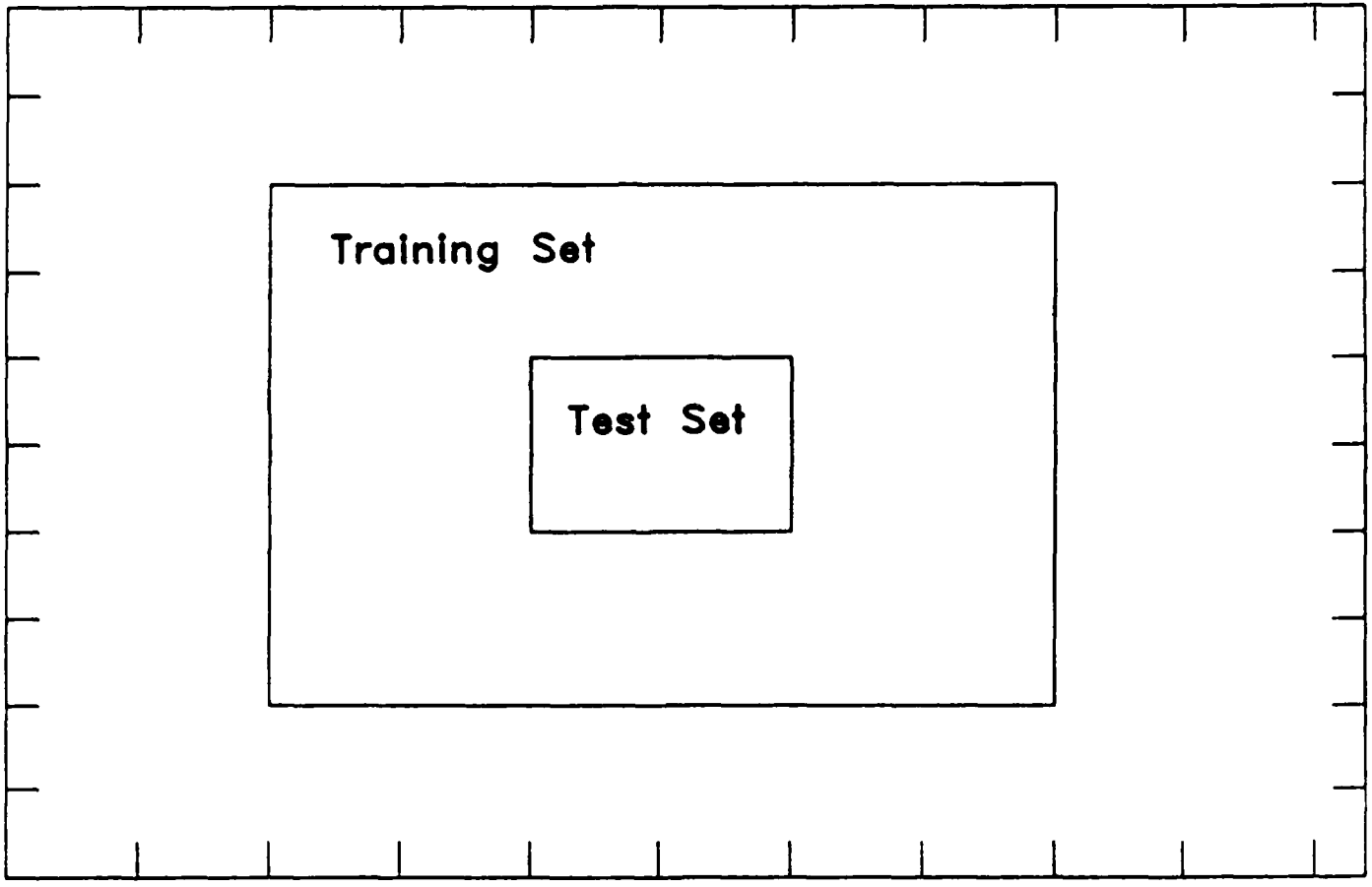

Fig. 7

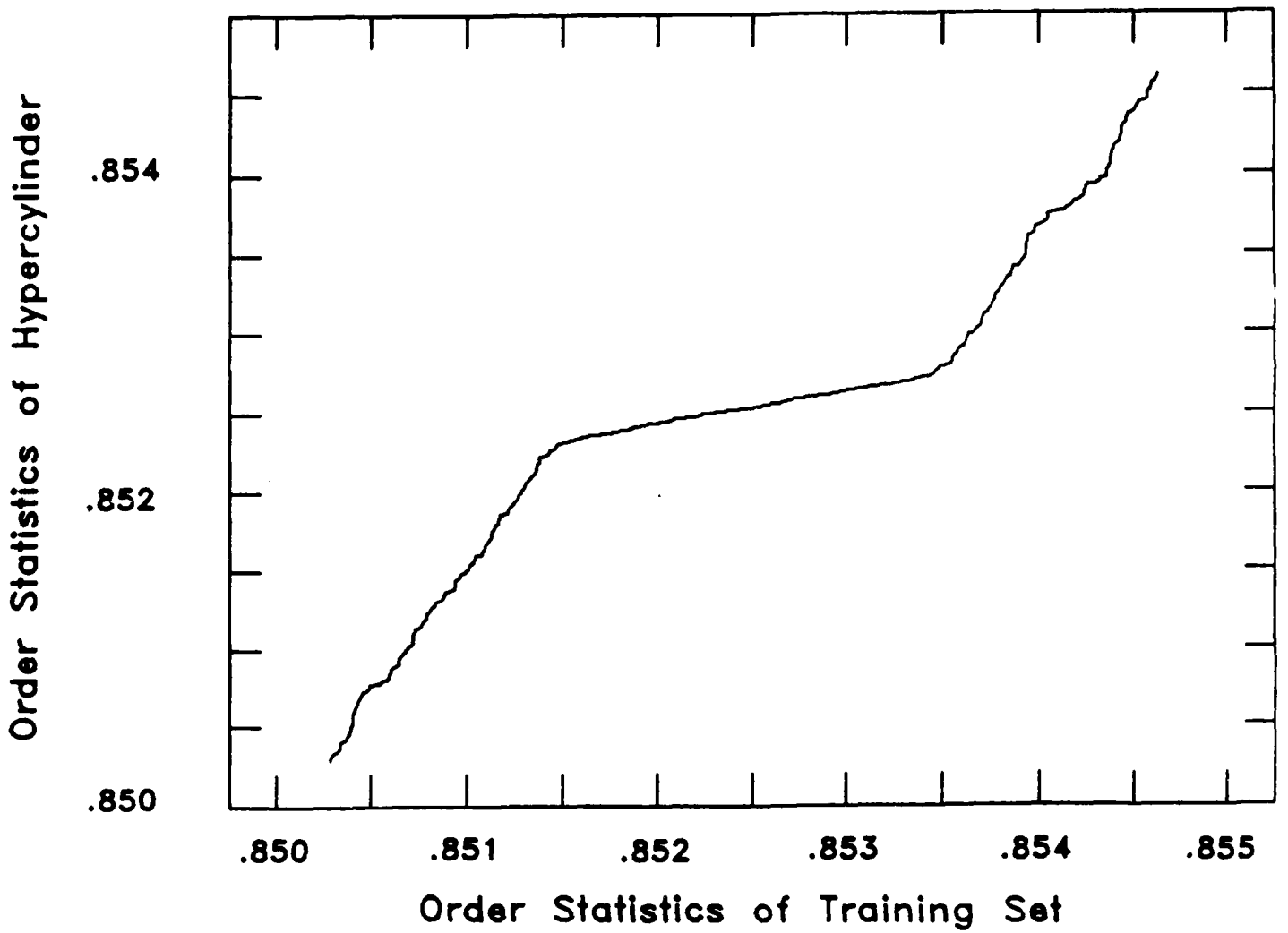
Order Statistics of Hypercylinder
.854
.852
.850
.850
.851
.852
.853
.854
.855
Order Statistics of Training Set

Fig. 8

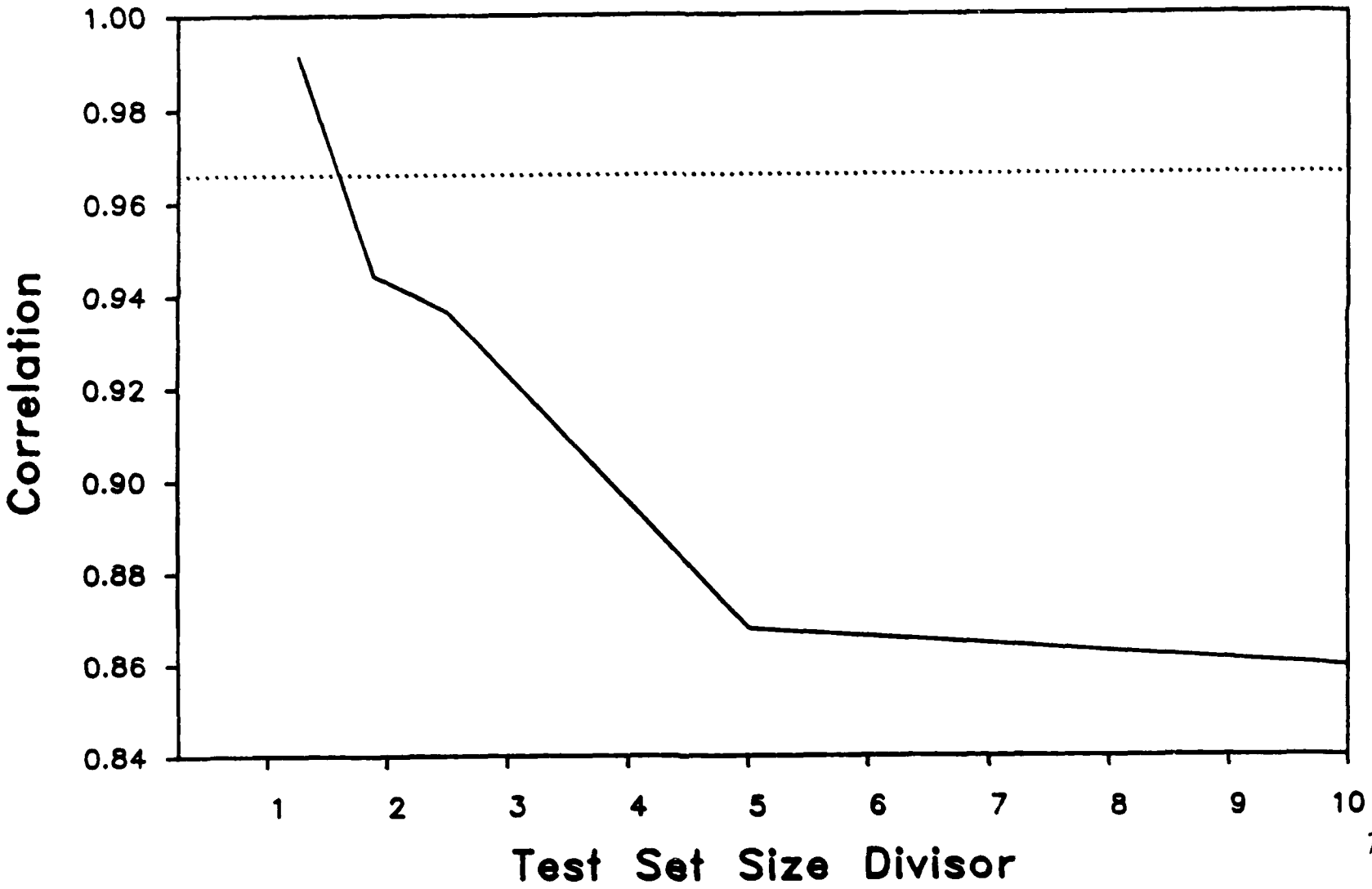
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Correlation
1
2
3
4
5
6
7
8
9
10
Test Set Size Divisor

Fig. 9

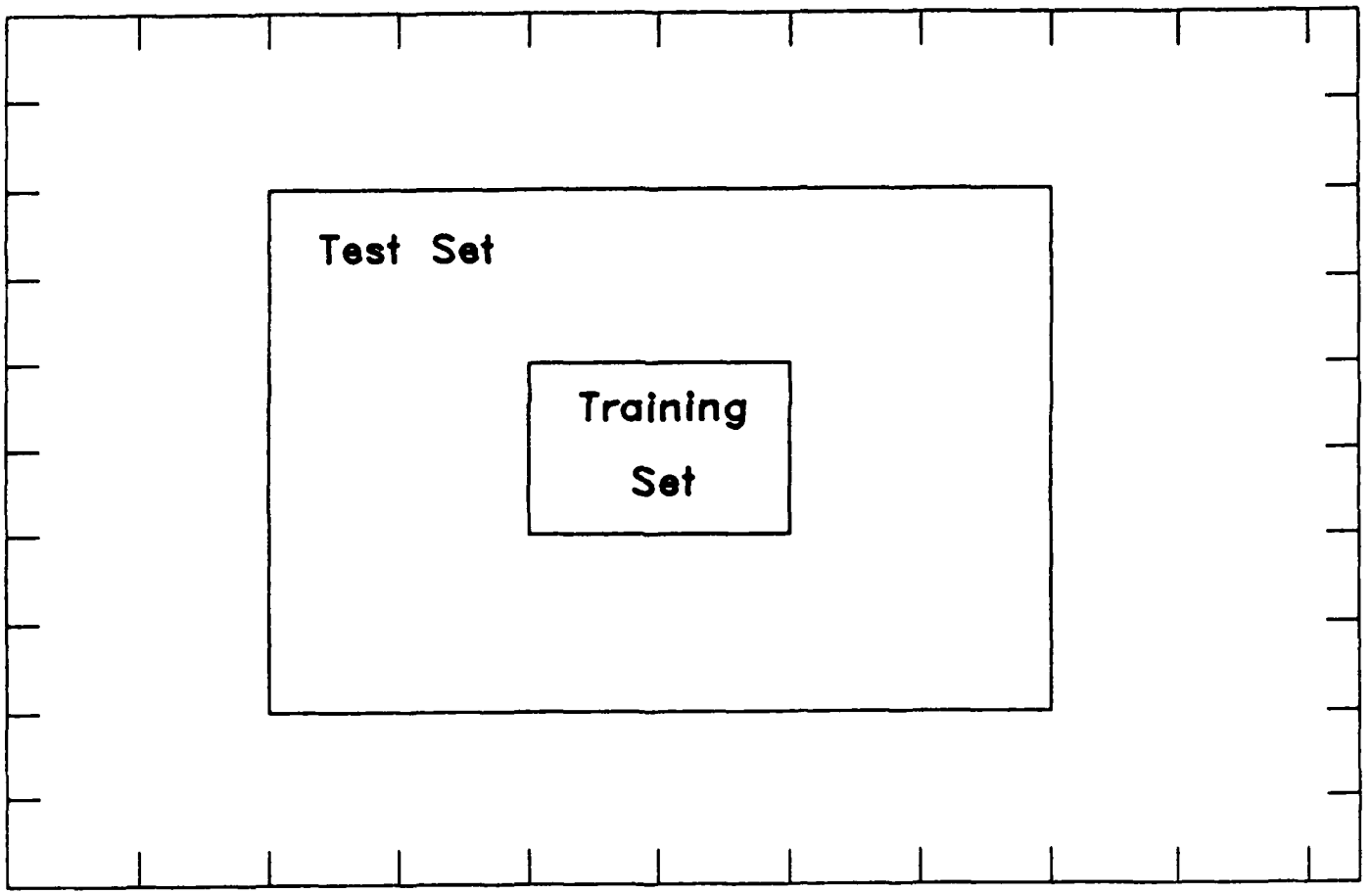

Fig. 10

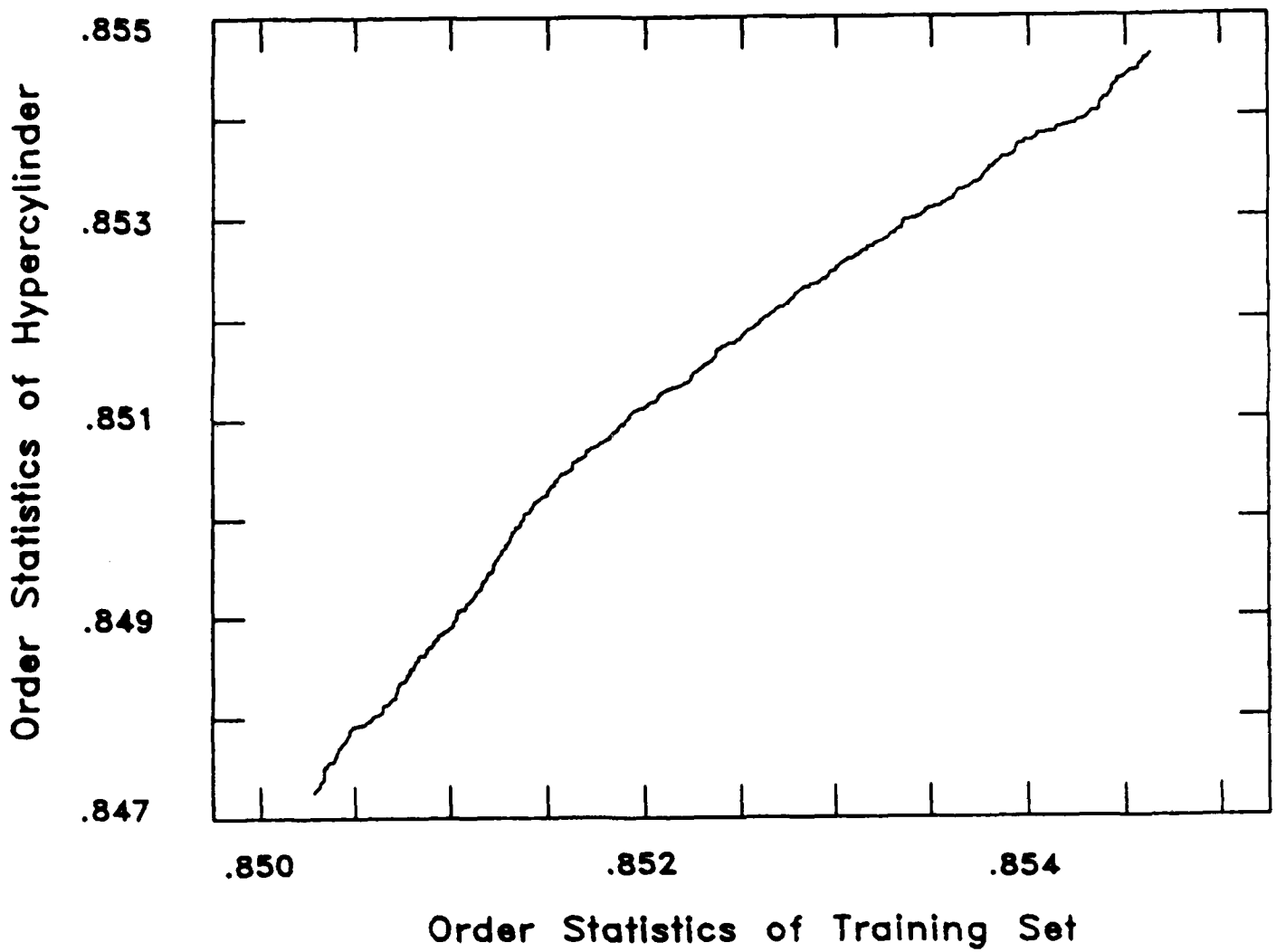
.855
.853
.851
.849
.847
Order Statistics of Hypercylinder
.850
.852
.854
Order Statistics of Training Set

Fig. 11

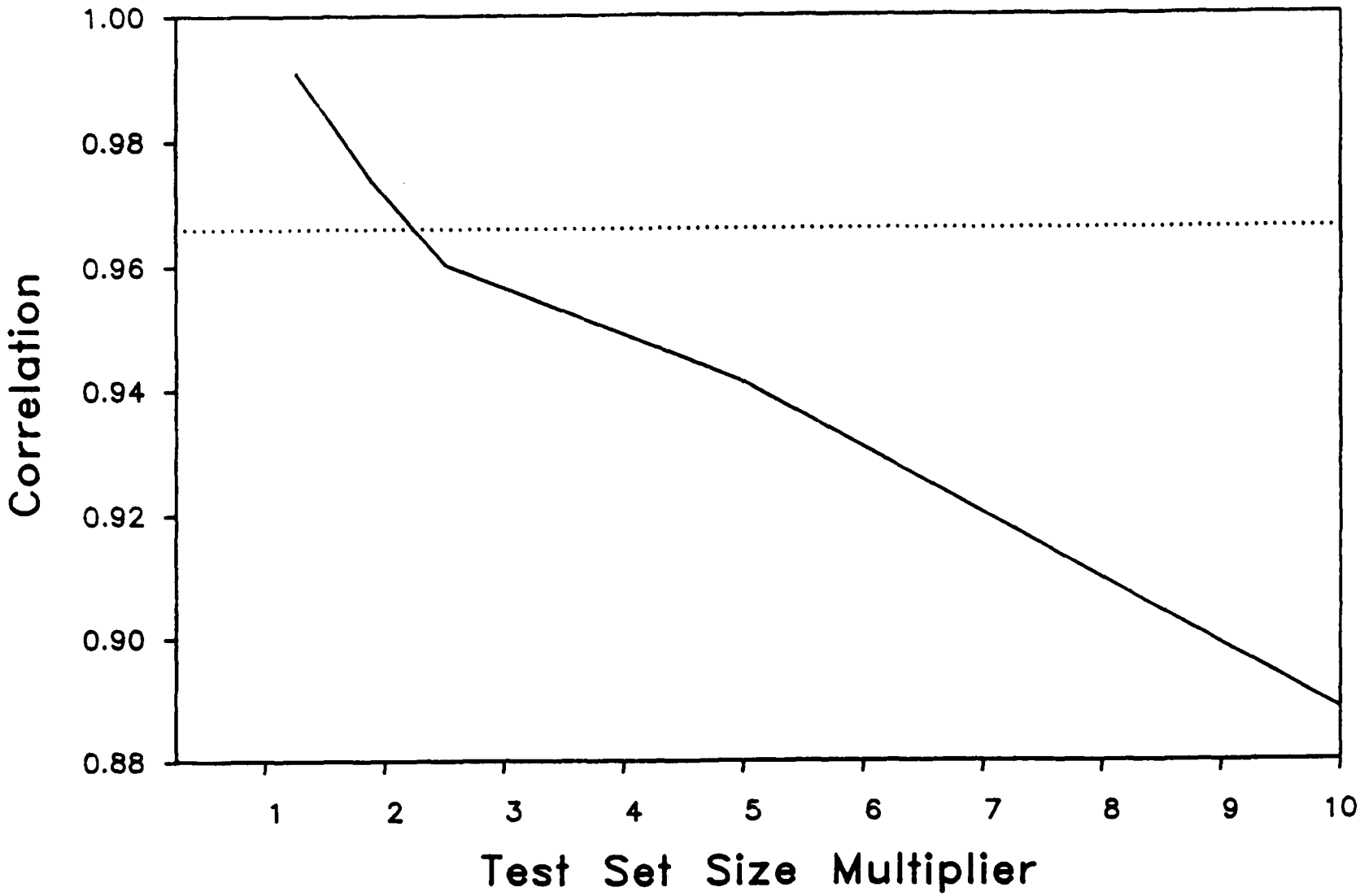
1.00
0.98
0.96
0.94
0.92
0.90
0.88
Correlation
1
2
3
4
5
6
7
8
9
10
Test Set Size Multiplier

Fig. 12

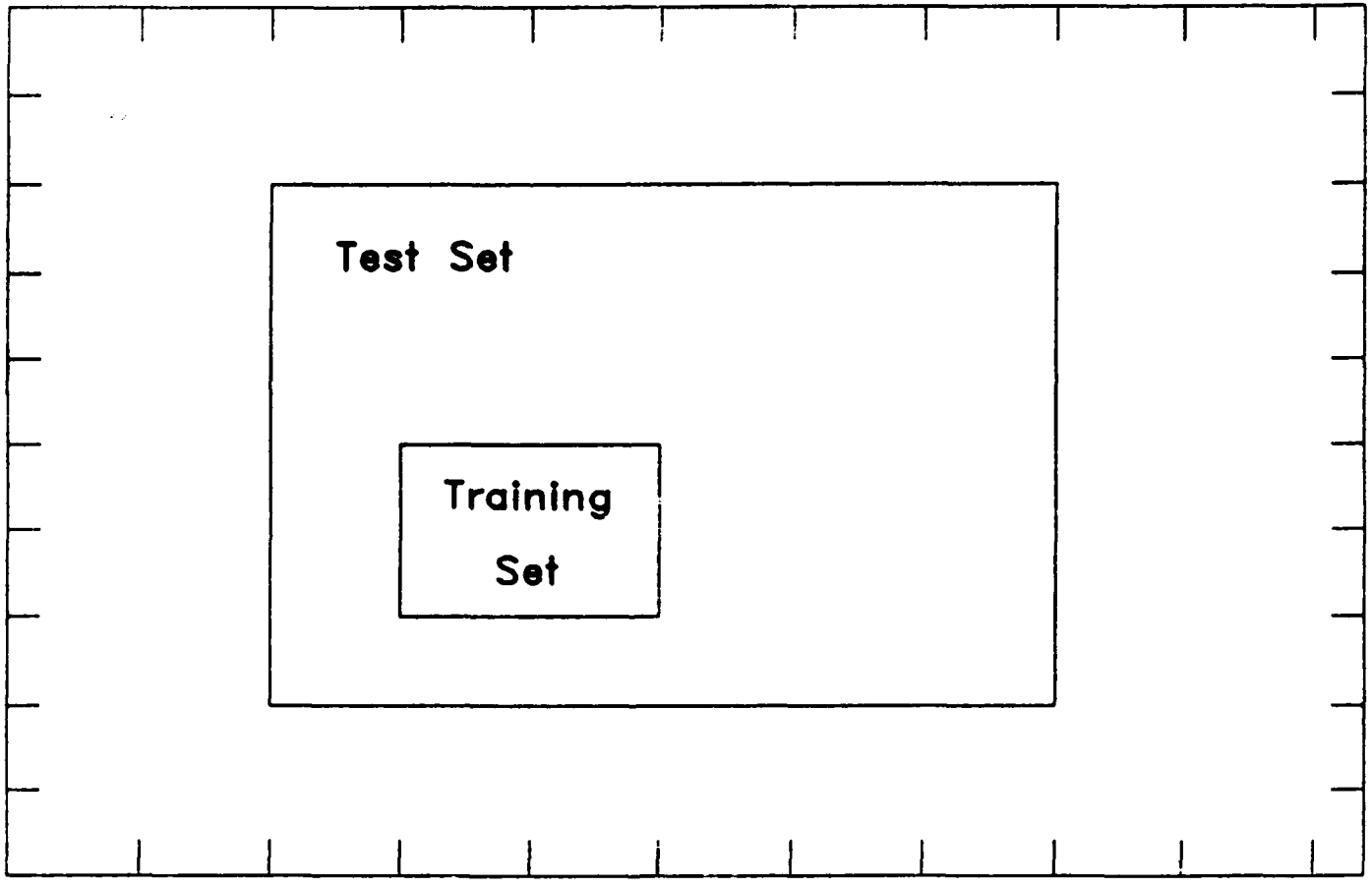

Fig 13

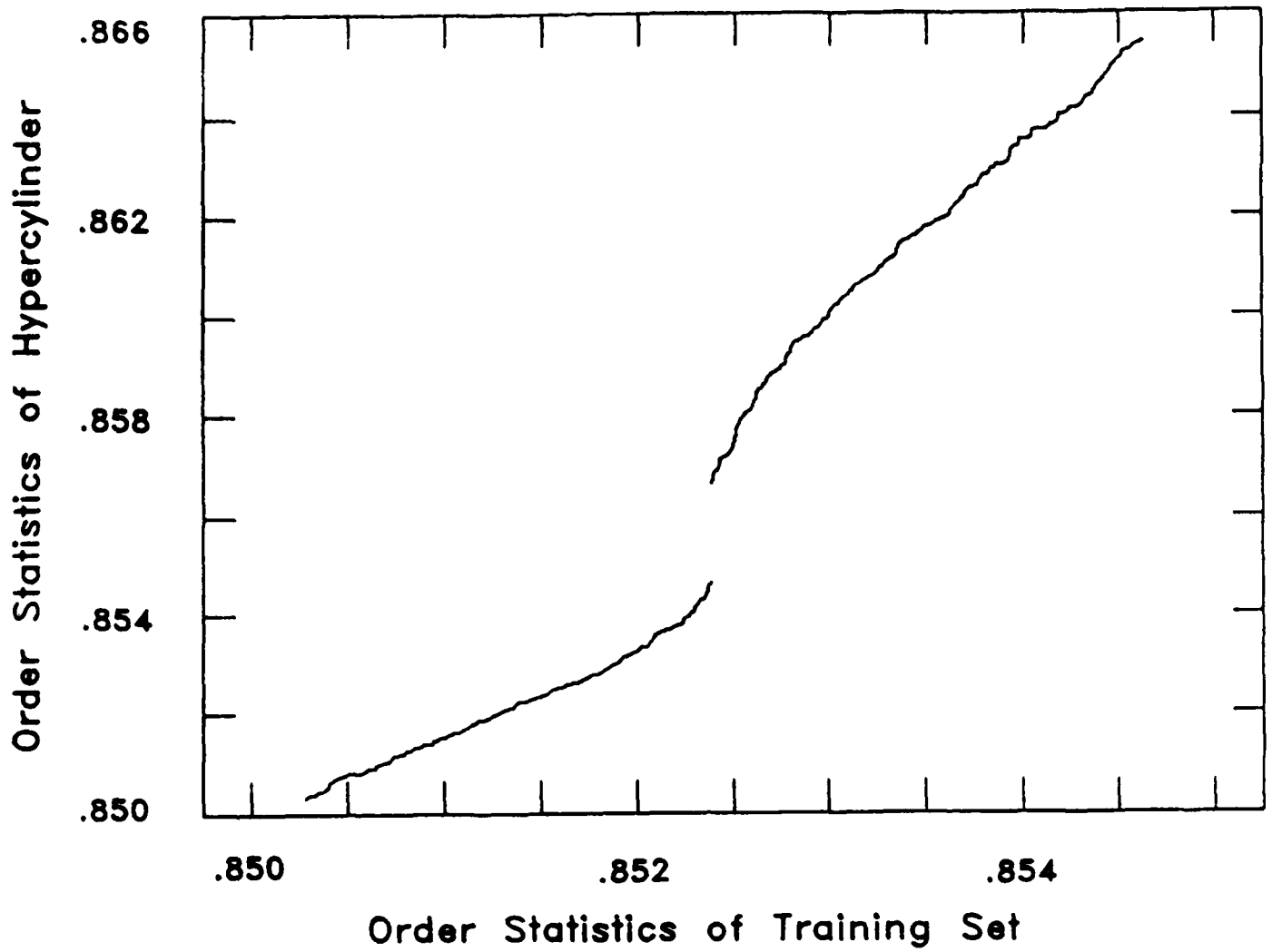
.866
.862
.858
.854
.850
Order Statistics of Hypercylinder
.850
.852
.854
Order Statistics of Training Set

Fig. 16

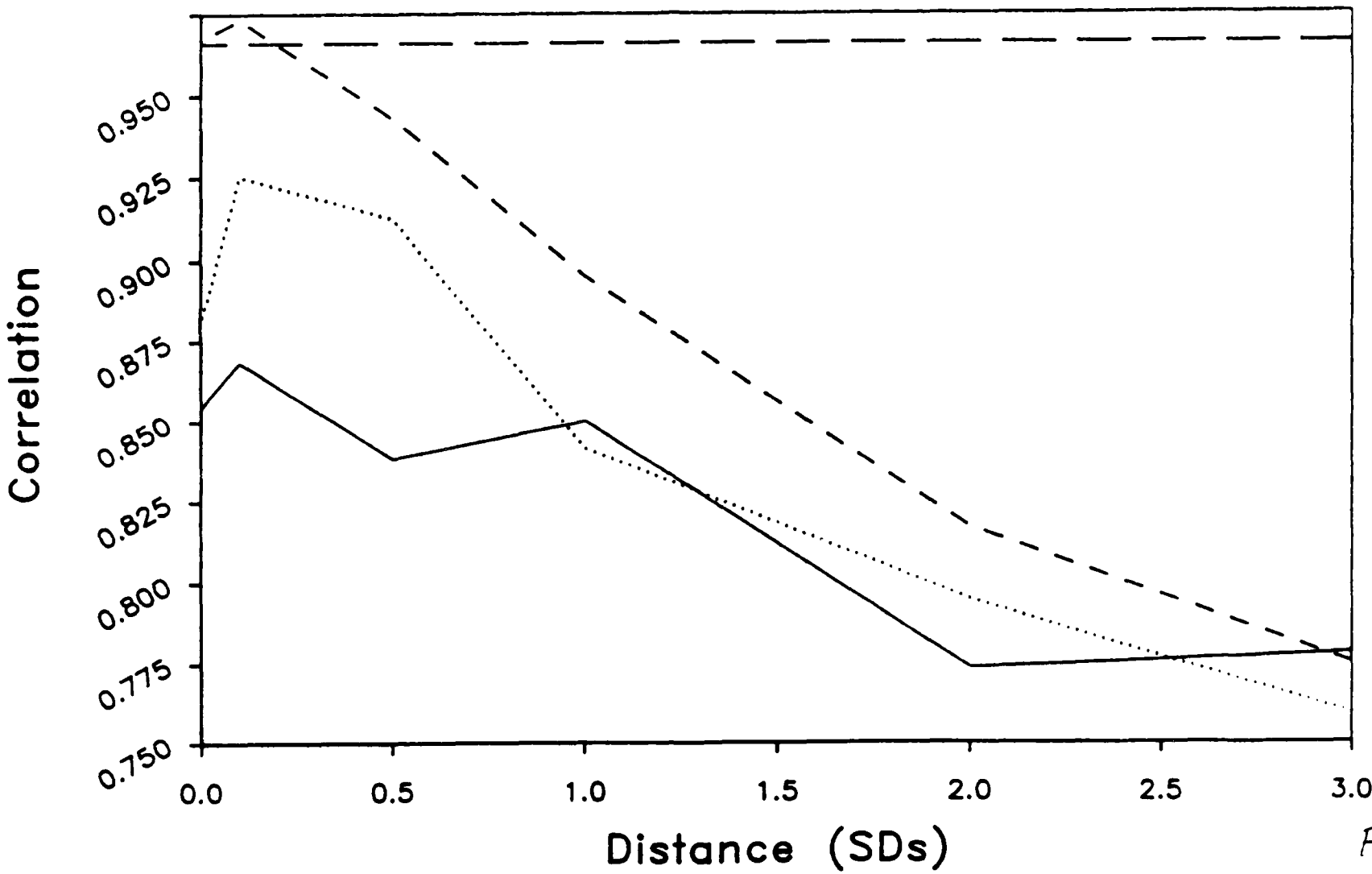
Correlation
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Distance (SDs)

Fig. 14

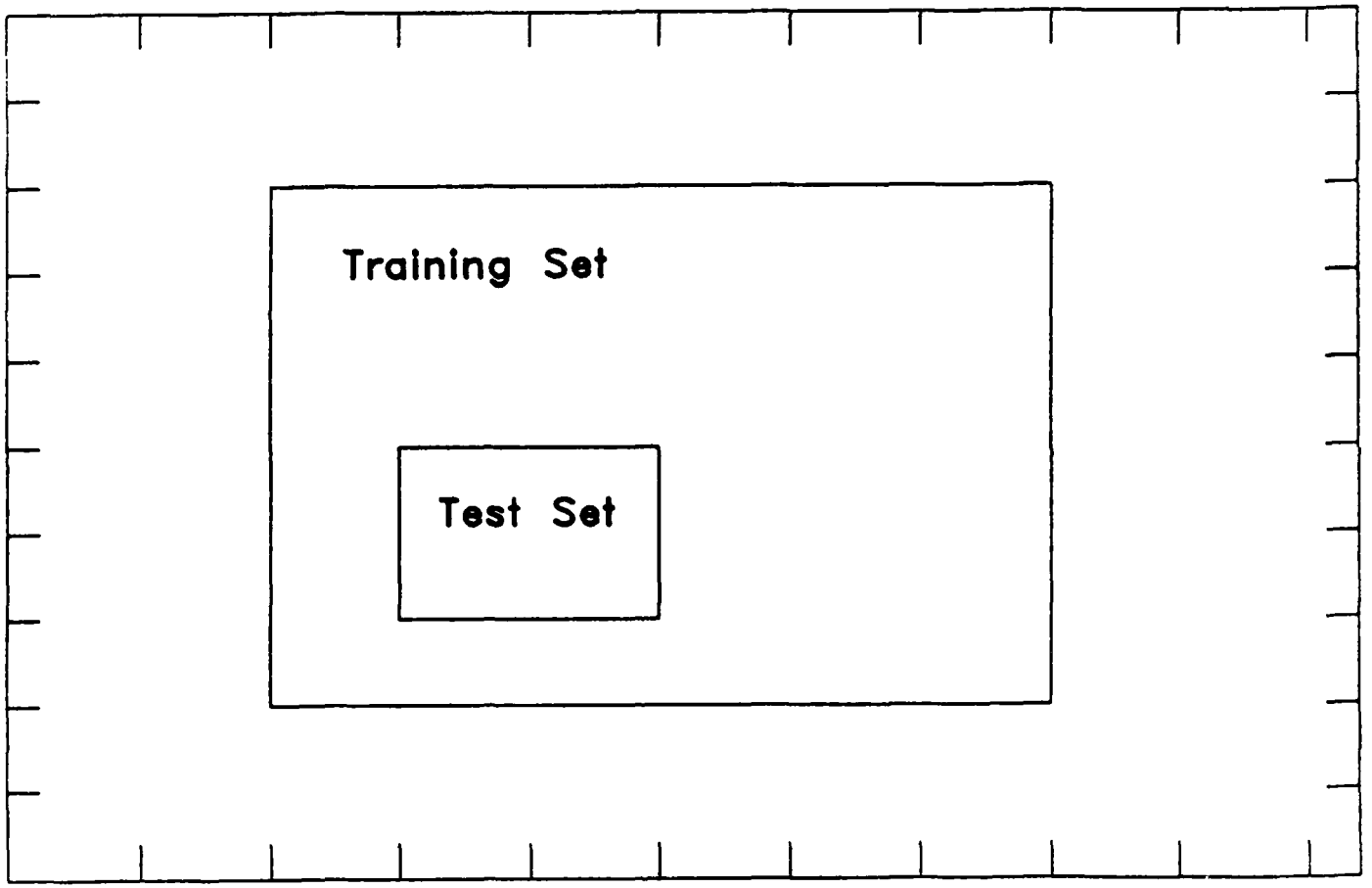

Fig. 6

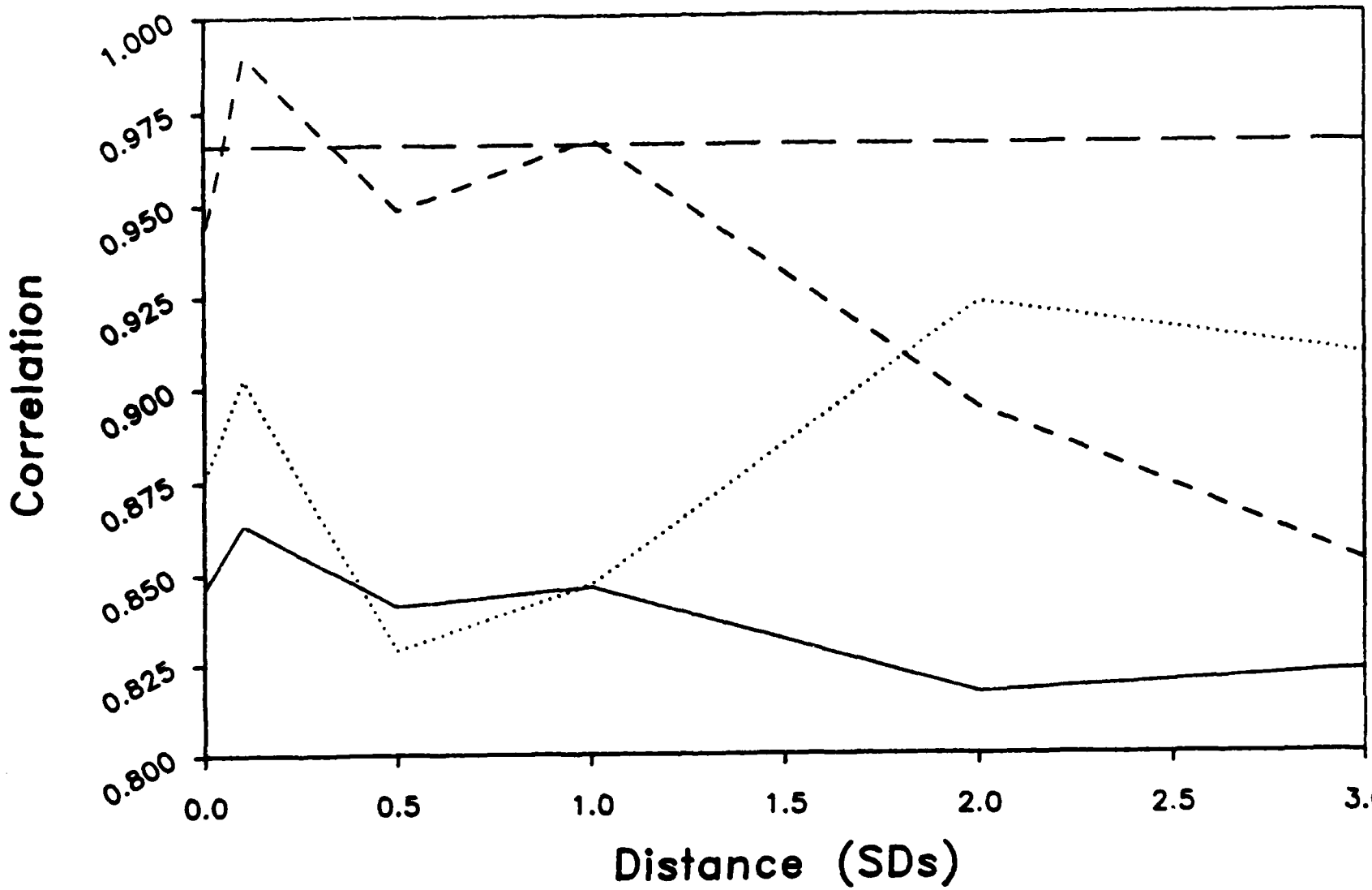
Correlation
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Distance (SDs)

Fig. 17

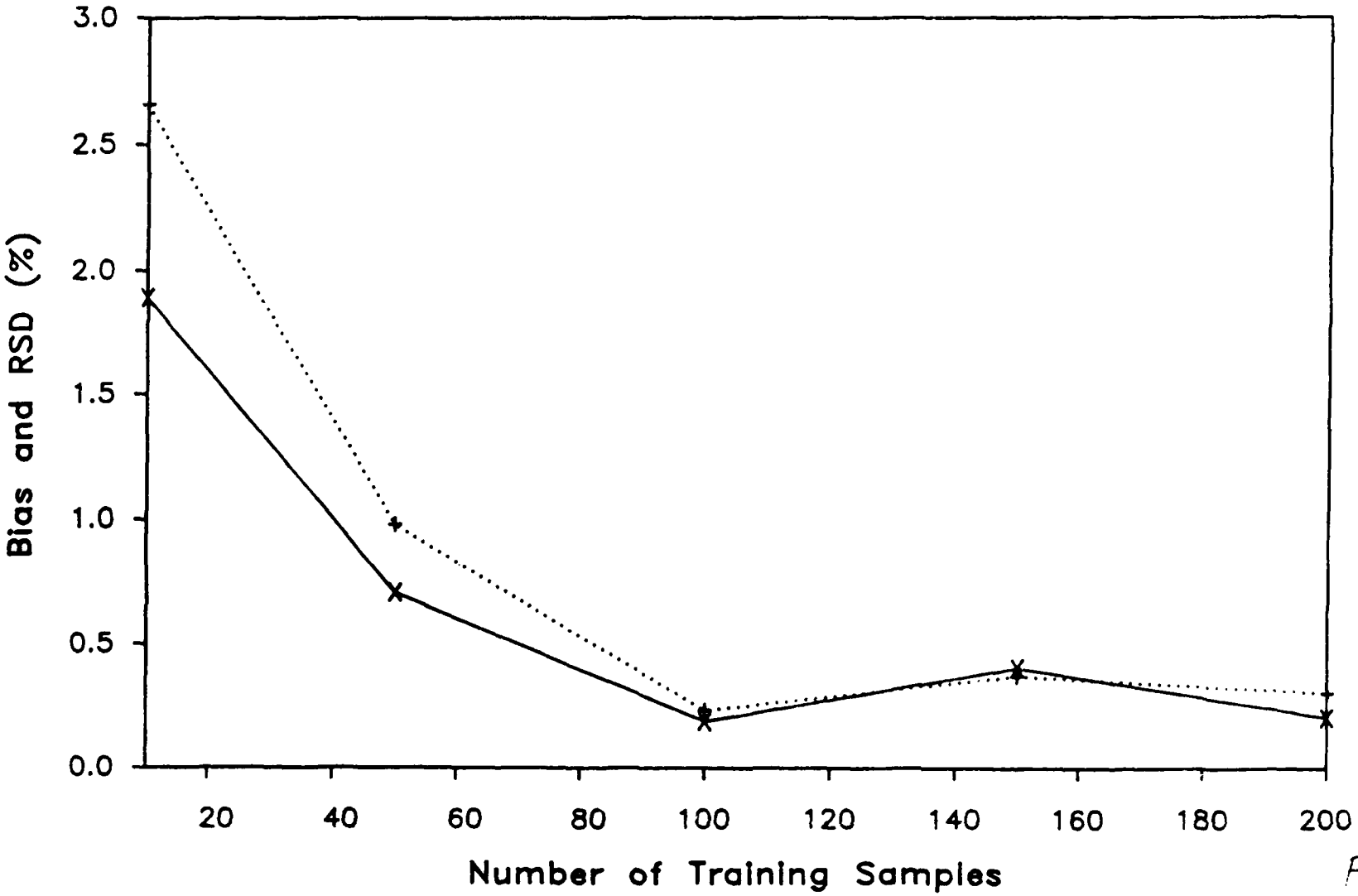
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Bias and RSD (%)
20
40
60
80
100
120
140
160
180
200
Number of Training Samples

Fig. 18

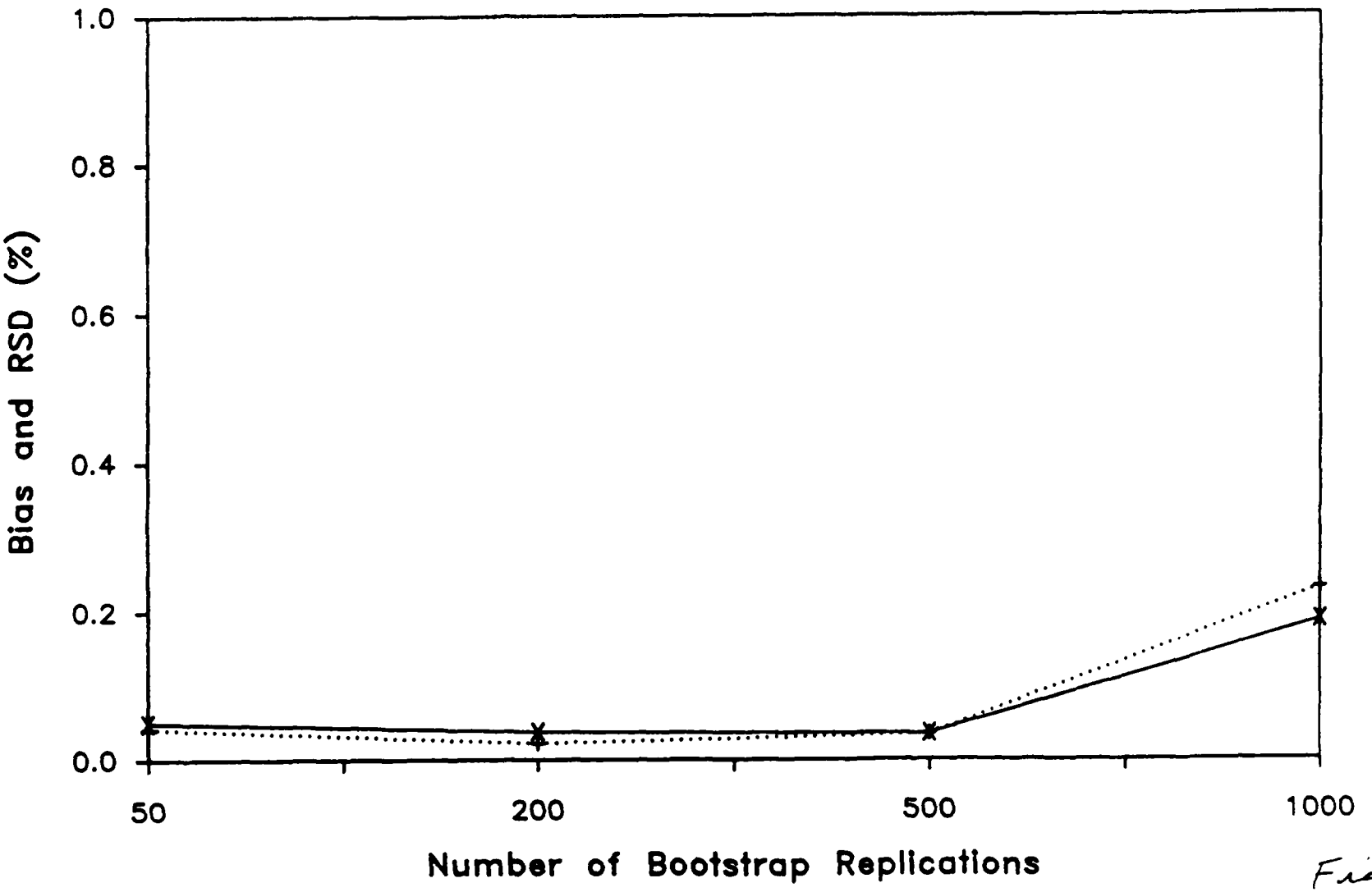
1.0
0.8
0.6
0.4
0.2
0.0
Bias and RSD (%)
50
200
500
1000
Number of Bootstrap Replications

Fig 19.

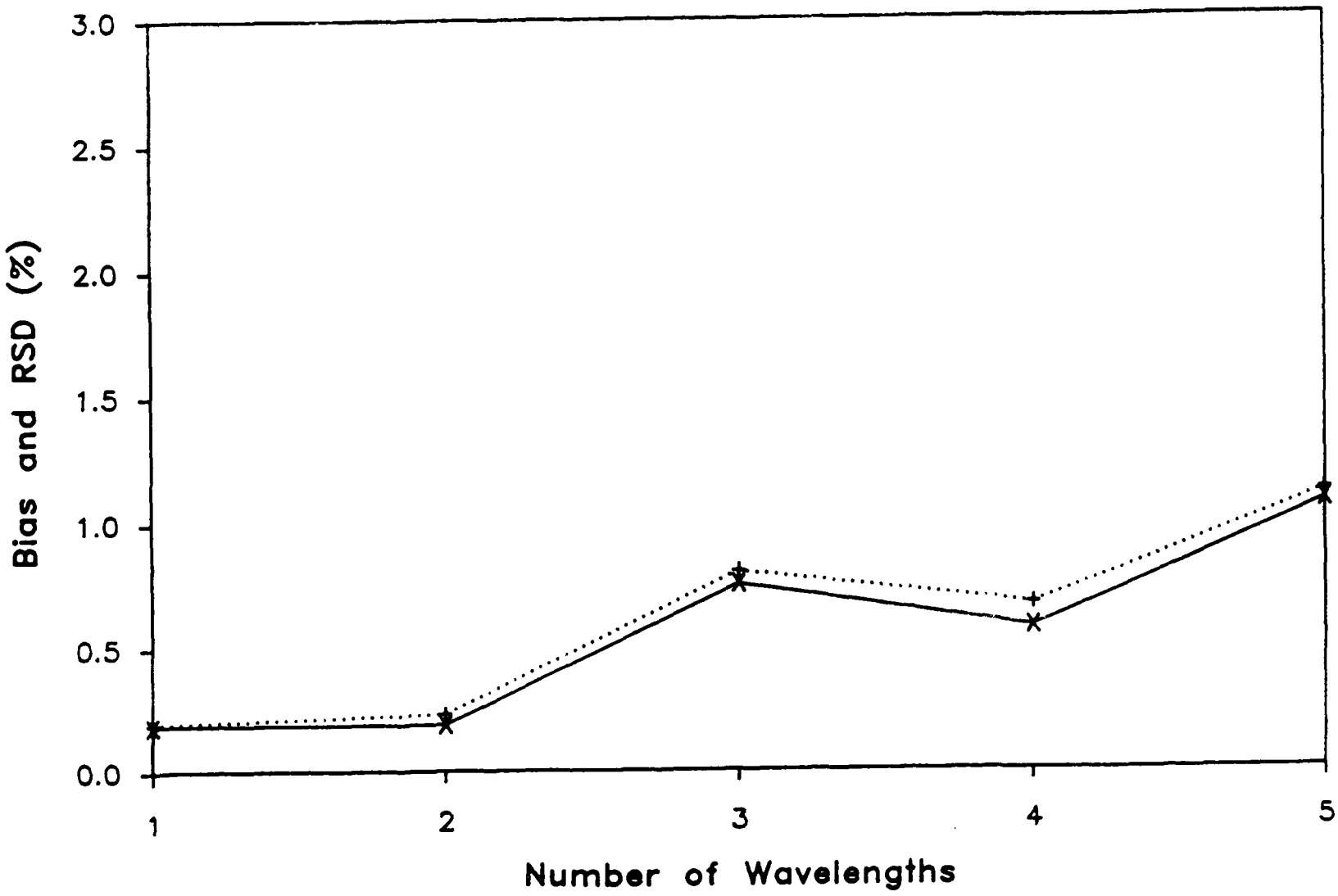

Fig. 20

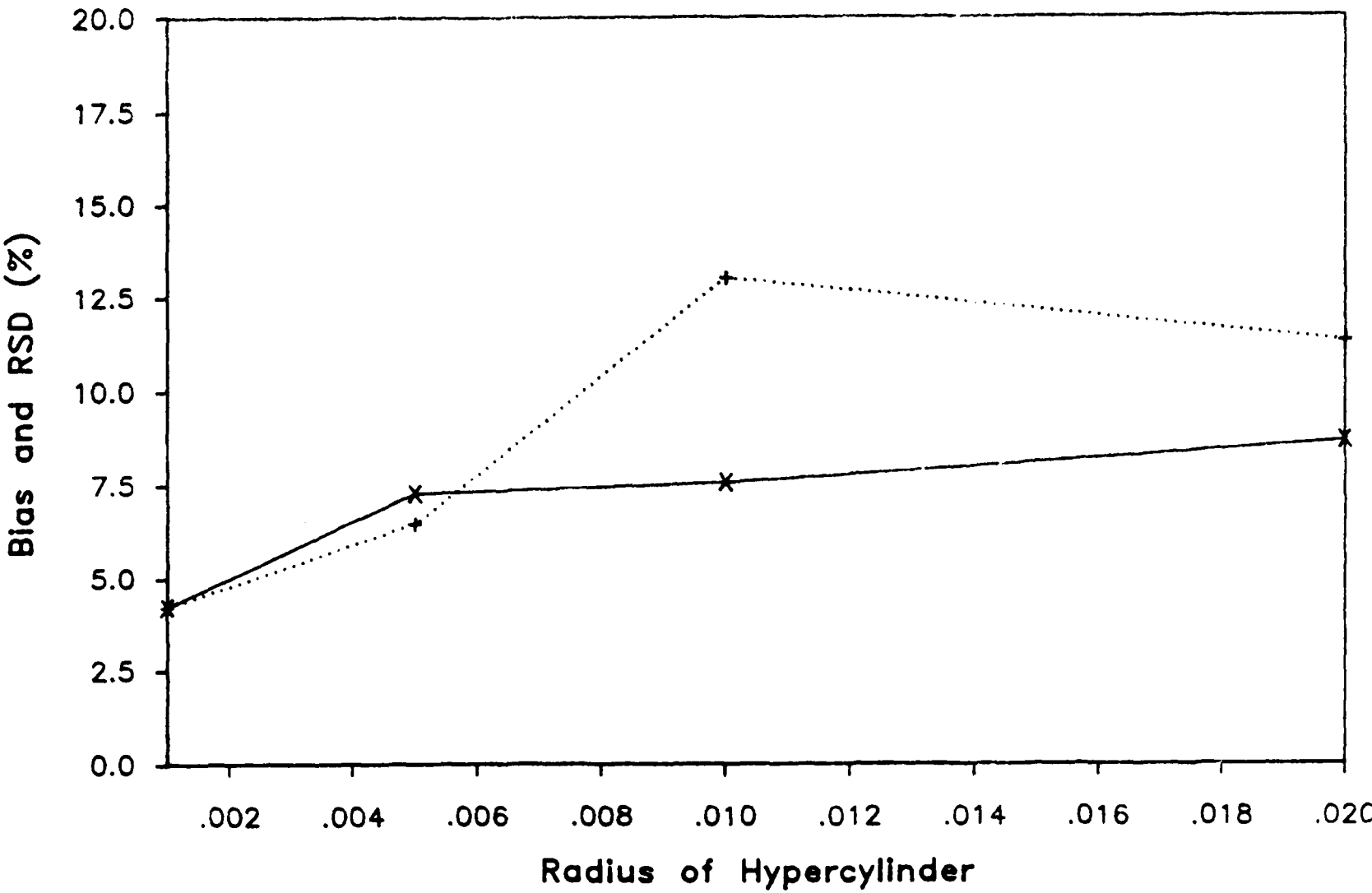
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Bias and RSD (%)
.002
.004
.006
.008
.010
.012
.014
.016
.018
.020
Radius of Hypercylinder

Fig. 2

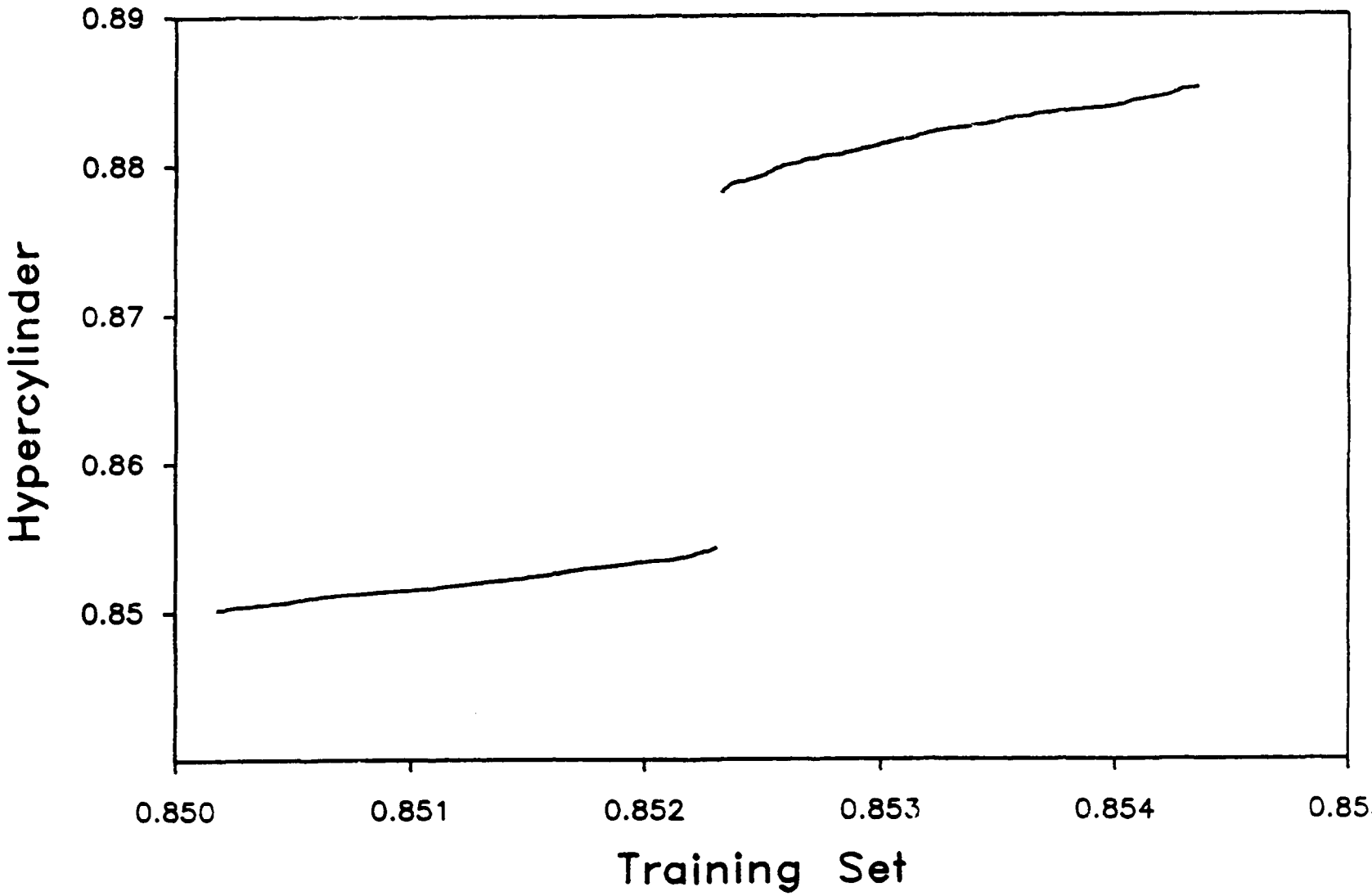
0.89
0.88
0.87
0.86
0.85
Hypercylinder
0.850
0.851
0.852
0.853
0.854
0.855
Training Set

Fig. 22

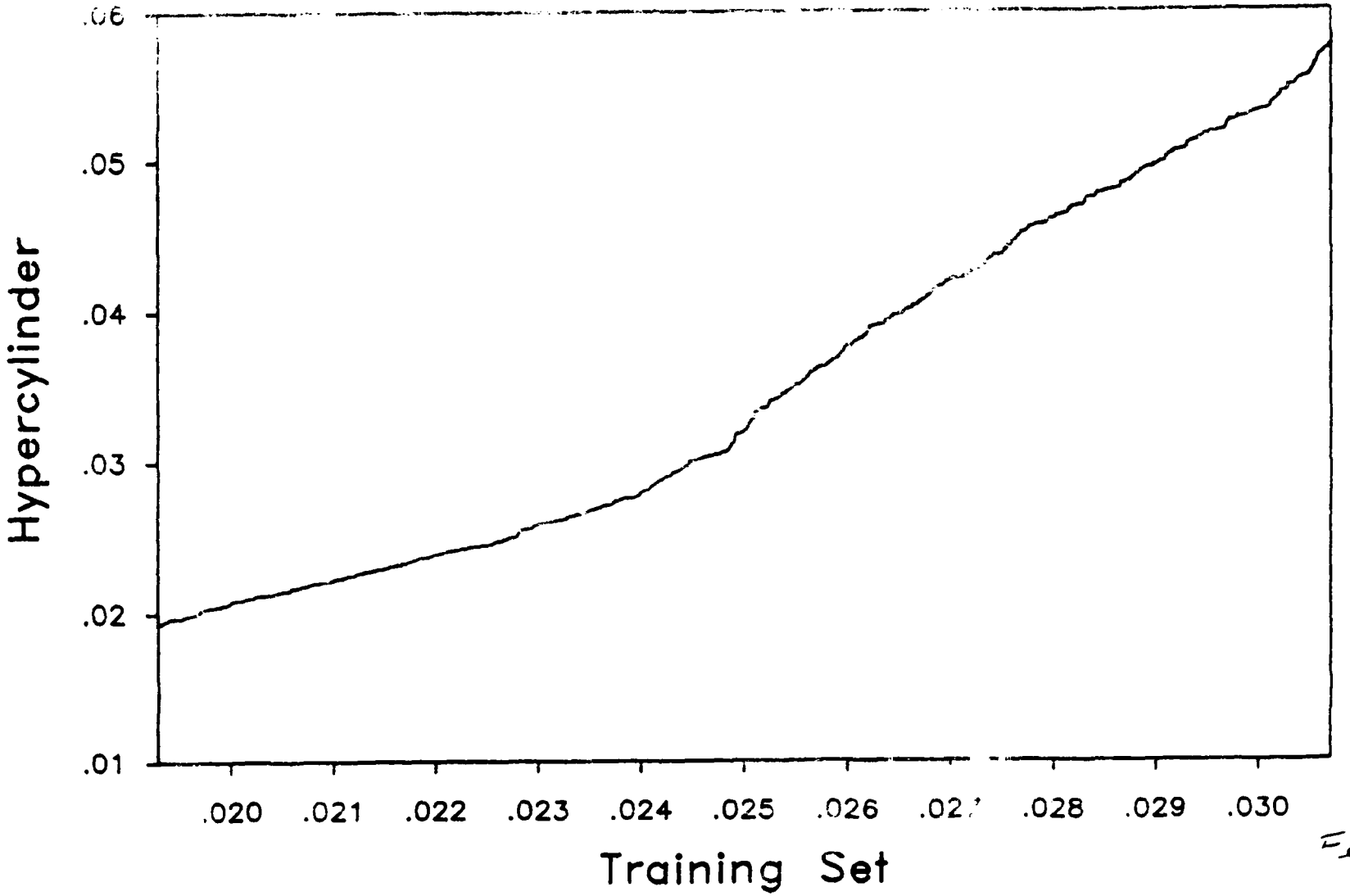
.06
.05
.04
.03
.02
.01
Hypercylinder
.020
.021
.022
.023
.024
.025
.026
.028
.029
.030
Training Set

Fig. 23

LIST OF SYMBOLS

Special defined operations:

$\sim$	random number on $0<x<1$; Monte Carlo integration of a continuous uniform distribution
$\sim(\cdot)$	creates a bootstrap distribution containing m elements for a set of real samples, and finds the center of this bootstrap distribution
$[x]$	the greatest-integer function of a scalar, matrix, or array
$\}(\cdot)$	ordered elements of x (x is a matrix or array)
$=$	equals, or "is replaced by" when the same variable appears on both sides of =

Scalars:

n	the training-set, test-set, and validation-set size, i.e., the number of samples that the set contains

d	the number of wavelengths and the dimensionality of the analytical space
m	the number of sample-set replications forming a bootstrap distribution (user-determined)
i	an index for counting rows in a matrix or array
j	an index for counting columns in a matrix or array
n_{λ}	the number of replicate spectral points falling inside a hypercylinder
p	proportion of a distance distribution to trim from each end of the distribution

Matrices, vectors, and arrays:

$B = (b_{ij})_{m,d}$	m-by-d bootstrap distribution of training-set sample spectra
$B_{(X)} = (b_{ij})_{m,d}$	bootstrap distribution of test-set sample spectra
$B_{(V)} = (b_{ij})_{m,d}$	bootstrap distribution of validation-set sample spectra

$C = (c_j)_d$	center of the bootstrap distribution B
$P = (p_{ij})_{m,n}$	training-set sample numbers selected for the bootstrap-sample sets used to calculate bootstrap distribution
$T = (t_{ij})_{n,d}$	training-set sample spectra
$X = (x_{ij})_{n,d}$	test-set sample spectra
$V = (v_{ij})_{n,d}$	validation-set sample spectra
$K = (k_j)_n$	training-set sample numbers selected for a particular bootstrap sample
$B_{(x)} = (b_{(x)ij})_{n,d}$	bootstrap sample set used to calculate single rows of a bootstrap distribution (B, $B_{(X)}$, or $B_{(V)}$)
$S_{(T)} = (s_{(T)i})_m$	Euclidean distances of training-set replicates from C, the center of the bootstrap distribution of the training set
$S_{(X)} = (s_{(X)i})_m$	Euclidean distances of test-set replicates from C
$S_{(V)} = (s_{(V)i})_m$	Euclidean distances of validation-set replicates from C

$P_{(T)} = (p_i)_{m-2pm}$ set of (m-2pm) indices used for trimming distance distributions

$C_{(t)} = (c_{(t)i})_{2m-4pm}$ cumulative distribution function (CDF) formed by the trimmed and ordered elements of the training-set bootstrap distribution; CDF has (2m-4pm) elements

$C_{(X)} = (c_{(X)i})_{2m-4pm}$ CDF formed by the trimmed and ordered elements of the test-set and training-set bootstrap distributions

$C_{(V)} = (c_{(V)i})_{2m-4pm}$ CDF formed by the trimmed and ordered elements of the validation-set and training-set bootstrap distributions

TECHNICAL REPORT DISTRIBUTION LIST, GEN

	No. Copies
Office of Naval Research Attn: Code 1113 800 N. Quincy Street Arlington, Virginia 22217-5000	2
Dr. Bernard Douda Naval Weapons Support Center Code 50C Crane, Indiana 47522-5050	1
Naval Civil Engineering Laboratory Attn: Dr. R. W. Drisko, Code L52 Port Hueneme, California 93401	1
Defense Technical Information Center Building 5, Cameron Station Alexandria, Virginia 22314	12 high quality
DTNSRDC Attn: Dr. H. Singerman Applied Chemistry Division Annapolis, Maryland 21401	1
Dr. William Tolles Superintendent Chemistry Division, Code 6100 Naval Research Laboratory Washington, D.C. 20375-5000	1
Dr. David Young Code 334 NORDA NSTL, Mississippi 39529	1
Naval Weapons Center Attn: Dr. Ron Atkins Chemistry Division China Lake, California 93555	1
Scientific Advisor Commandant of the Marine Corps Code RD-1 Washington, D.C. 20380	1
U.S. Army Research Office Attn: CRD-AA-IP P.O. Box 12211 Research Triangle Park, NC 27709	1
Mr. John Boyle Materials Branch Naval Ship Engineering Center Philadelphia, Pennsylvania 19112	1
Naval Ocean Systems Center Attn: Dr. S. Yamamoto Marine Sciences Division San Diego, California 91232	1